Microbial Evolution:
Robustness and Innovation in Metabolic Networks and the Evolution of Terminal Cell Differentiation

Dissertation

zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
João Frederico Matias Rodrigues
aus
Portugal

Promotionskomitee:
Prof. Dr. Andreas Wagner (Vorsitz)
Prof. Dr. Olivier Martin
Prof. Dr. Homayoun C. Bagheri

Zürich, 2011

Abstract

Microbes include some of the most ancient and ubiquitous organisms alive today. They are the class of organisms that have most impact on human society. They are involved in major geophysical cycles, used in biotechnology and are one of the main causes of disease in humans. Despite their importance, recent developments have shown that the microbial world remains largely uncharacterized. Advances in technology have allowed a glimpse at the phylogenetic relationships of microbes and made possible the systematic study of microbial evolution. Many recent findings have already forced the revision of our understanding of evolution in this context. In my thesis I have focused on two topics concerning microbial evolution. The first topic regards the evolution of metabolic networks, the networks of chemical reactions occurring in organisms. I investigate this topic by simulating the evolution of metabolic networks under selection for viability in a given environment. Random metabolic networks with the same phenotype produced through this process share a core of super-essential reactions, reactions essential to all produced networks. My results show that it is possible through single mutations to arrive at a metabolic network very different from the initial metabolic network. This property found in the genotype-phenotype maps of metabolic networks indicates that such networks can potentially access many novel phenotypes than would otherwise be possible. The second topic in my thesis regards the evolution of terminal and reversible differentiation in multicellular cyanobacteria. Reversibly differentiated cells in some organisms enable them to regenerate or reproduce through fragmentation. However, despite the potential benefits of reversibly differentiated cells, cells in many organisms are instead terminally differentiated, and are therefore unable to produce other cell types. In my thesis, I have explored the conditions driving the evolution of terminal and reversible differentiation. The results show that although both cell interaction topology and differentiation costs play a role, differential cell growth between cell types is the main factor controlling the type of differentiation evolving. I find that the cell type that becomes the germline is the fastest growing.

Zusammenfassung

Zu den Mikroben gehören einige der ältesten und ubiquitären Organismen die heute leben. Sie sind die Klasse von Organismen, die den grössten Einfluss auf die menschliche Gesellschaft haben. Sie sind zum Beispiel in den wichtigsten geophysikalischen Zyklen beteiligt, werden in der Biotechnologie verwendet und sind eine der Hauptursachen von Krankheiten beim Menschen. Trotz ihrer grossen Bedeutung zeigt die jüngste Entwicklung, dass die mikrobielle Welt weitgehend noch nicht charakterisiert ist. Technologische Fortschritte erlauben einen Einblick in die phylogenetischen Beziehungen der Mikroben und ermöglichen somit die systematische Untersuchung der mikrobiellen Evolution. Viele neue Erkenntnisse haben uns bereits dazu gezwungen, unser Verständniss dieser Evolution zu revidieren. Meine Dissertation konzentriert sich auf zwei Themen der mikrobiellen Evolution. Das erste Thema betrifft die Entwicklung von metabolischen Netzwerken, Netzwerke, welche die chemischen Reaktionen in einem Organismus darstellen. Dies erreiche ich durch die Simulation der Evolution von metabolischen Netzwerken unter Selektion für die Lebensfähigkeit in einer bestimmten Umgebung. Hierbei werden zufällige metabolische Netzwerke mit dem gleichen Phänotyp erzeugt, und es zeigt sich, dass alle diese Netzwerke die gleichen sehr wichtigen Reaktionen besitzen. Diese Ergebnisse zeigen, dass es möglich ist durch einzelne Mutationen zu einem metabolischen Netzwerk zu gelangen, welches sich sehr vom Ausgangsnetzwerk unterscheidet. Diese Eigenschaft, die in Genotyp-Phänotyp-Karten metabolischer Netzwerke gefunden wurde, zeigt, dass solche Netzwerke viel mehr neue Phänotypen finden als es sonst möglich wäre.
Das zweite Thema meiner Arbeit bezieht sich auf die Entwicklung von irreversibler und reversibler Zelldifferenzierung in mehrzelligen Cyanobakterien. In einigen Organismen erlauben reversibel differenzierte Zellen, dass sich die Zellen regenerieren oder durch Fragmentierung reproduzieren. Trotz der möglichen Vorteile von reversibel differenzierten Zellen, sind Zellen in vielen Organismen irreversibel differenziert und haben daher nicht die Möglichkeit andere Zelltypen zu erzeugen. In meiner Arbeit habe ich die Bedingungen untersucht, welche die Evolution von reversibler und irreversibler Zelldifferenzierung beeinflussen. Die Ergebnisse zeigen, dass, obwohl Zellinteraktionstopologie und Differenzierungskosten eine Rolle spielen, unterschiedliches Zellwachstum zwischen Zelltypen der wichtigste Faktor ist, der die Art der Differenzierung kontrolliert. Ich stelle fest, dass der Zelltyp welcher zur Keimzelle wird, der am schnellsten wachsende ist.

Acknowledgments

I dedicate this thesis to my wife Hanna and son Francisco, my parents Dulce and Vitor, and my siblings Sandra and Miguel. I thank my supervisor Andreas for giving me the opportunity to work in this exciting field and learn much in the process. A special thanks goes out to Homayoun Bagheri, Olivier Martin, Jorge Pacheco, Areejit Samal, Daniel Rankin, and Francisco Santos whose help was instrumental for me to accomplish this thesis. Many thanks to Christiane, Annette, Rudolf, Karthik, Carlos, Evandro, Manuel, Elias, Giovanni, Riddhiman, Eric, Bing and Sorcha with whom I shared my day to days and who made it fun to come to work. Thanks to Nicole for helping me with the thesis formatting and with the german abstract. Last, but not least, big thanks to Jeremiah, Saurabh and Aditya with whom I shared my office or most of my time and who undoubtedly gave me many ideas and improved my understanding of biology.

Contents

1 Introduction

Almost four centuries since the discovery of microbes, they are to this day the source of many scientific surprises. One of the most ancient and diverse life forms on earth, they are found to colonize almost every corner of our planet from glaciers to hot springs, acid mine drainages, deserts. They occur deep in the earth, in the oceans, and even in our body. Single microbes are invisible to the naked eye, but their impact in our world is hard to ignore. They are involved in most major geophysical processes and cycles [1, 2]. In the oceans, microbes in phytoplankton are responsible for almost half of the world's carbon fixation [3]. Their importance for human society encompasses many fields, from medicine to biotechnology and food production. They are also the cause of many human diseases.

1.1 The vast unknown microbial world

For a long time, the characterization of microbes required the cultivation and isolation of single species in the laboratory. In this manner, more than 5000 microbes have been characterized until now [4]. However, in the past decades it has become apparent that a large number of microbes are not easily cultured in the lab. The exact number of unculturable species remained elusive until recently. The development of methods that allowed the study of uncultured microbes has made it evident that less than 1% of all living microbes have been cultured so far [4]. One problem in estimating the diversity of microbial species is the lack of a single definition of species. In practive, most culture independent studies define species based on a threshold of sequence identity. Nonetheless, microbial diversity is so large that just on the human body 17'000 species can be found. On human skin 500 species have been found [5], 500 in the oral cavity [6, 7] and 16'000 in the gut [8]. Additionally, there are an estimated 800'000 insect species, of which at least 10% may carry obligate symbionts [4]. Similarly large numbers of microbial species can be found in samples taken in diverse environments from the Sargasso sea (where 1800 species were found) [9] to different soil samples (where 847 species were found) [10]. With more than 99% of the microbial world being unknown, it is a safe assumption that many more surprises await discovery. Just as horizontal gene transfer has forced us to change our views of an evolution based on vertical transfer, what other surprises can we expect? The answer to this question

will lie in the experimental characterization of this vast and unexplored world. However, theoretical expectations have always been the guide for experiments. In my thesis, I attempt to delineate some of the expectations that can be made about the evolution of microbes by concentrating on two problems of the microbial world. The first concerns the evolution of genome-scale metabolic networks and is discussed in Chapters 2 and 3. The second concerns the evolution of terminal cell differentiation and is discussed in Chapter 4.

1.2 Microbial evolution

Before the discovery of molecular methods, our understanding of the genealogical relationships of microbes, and indeed of all other organisms, was based solely on shared morphological features. A morphological feature that arises in one organism can be expected, often but not always, to be present in its descendants. This argument has allowed biologists to group together organisms by morphological features, and obtain a first glimpse at the tree of life. The tree that describes all genealogical relationships between organisms up to the last common ancestor. Such a genealogical tree is essential for our understanding of the origin and the events that led to the existence of the organisms alive today.

1.2.1 Phylogenetic tree of microbes

The development of molecular methods led to the realization that it is possible to estimate divergence times between organisms [11]. One theoretical basis for this comes partly from the neutral theory of molecular evolution which shows that neutral mutations can spread in a population and get fixed through neutral drift. Additionally, under constant population size and mutation rate, the number of fixation of neutral mutants occurs at a constant rate [12]. The fraction of mutations that are neutral has remained a topic of controversy, however even mutations that are slightly deleterious or beneficial will have a probability of fixation similar to that of a neutral mutation when selection pressure is low or population sizes are small. To estimate divergence times using molecular methods orthologous proteins (and later DNA molecules) from two organisms were compared. Proteins or DNA molecules with more differences indicated that the organisms had diverged a longer time in the past. Genetic comparison has become a standard method to estimate the genealogical relationships between organisms. A tree of genealogical relationships built using this method was named a phylogenetic tree. Different genes are found to evolve at different rates. Consequently, not all genes can be used to produce a phylogenetic tree for a group of organisms. Genes that evolve fast offer accurate estimates of divergence times for recently diverged species. However,

they cannot be used for species that diverged far in the past. Another problem of this approach is the requirement that all organisms have the same gene for which the phylogenetic tree is being produced. In higher order organisms cytochrome c was found to be a well conserved gene with a rate of divergence useful in the generation of phylogenetic trees [13]. However, microbial organisms are far more ancient than higher order organisms and therefore the use of cytochrome c in this case proved to be of limited value. Instead, for establishing the genealogical relationships in the microbial world, ribosomal RNAs (rRNAs) proved to be the best markers. These molecules were found to be universally distributed, to maintain their function, and to evolve at a far slower pace than most proteins. While in the case of higher order organisms the trees based on morphology were found to be mostly in agreement with the genetically based trees (phylogenetic trees), at the microbial level the morphological trees proved to be inaccurate and led to several major reorganizations in the classification of microbes [14]. One of these was the separation of the archaebacteria and bacteria into two different kingdoms. Another important finding for the understanding of early events in cellular evolution was that thermophilic bacteria and archaebacteria may be the most ancient organisms. This was crucial for the understanding of the evolution of metabolic pathways, a topic that will be discussed in section 1.3.1.

1.2.2 Horizontal gene transfer

The largest surprise was that evolution at the microbial level is incompatible with the standard model of evolution in which offspring inherit their genetic information only from their parents (vertical gene transfer). Instead, the evolution of microbes is dominated by vertical gene transfer, but is also strongly shaped by cases of horizontal gene transfer [15, 16, 17]. Horizontal gene transfer occurs when organisms incorporate genes from organisms in the same population, even from different species. Several observations have been made that have slowly built the case for the importance of horizontal gene transfer. Inconsistencies have been observed in the phylogenetic trees of genes in microbial genomes when compared with the phylogenetic trees of other genes or ribosomal RNAs. Antibiotic resistance has also been observed to evolve across microbial species due to horizontal gene transfer [17]. The percentage of genes in microbial genomes that have been acquired through horizontal gene transfer is currently estimated to range between 1.6 and 32.6 percent [15, 18]. Horizontal gene transfer can occur through a number of processes such as natural transformation, conjugation, or transduction involving the action of mobile genetic elements such as phages, retroviruses, or transposons [19, 20]. These mechanisms are illustrated in Figure 1.1. In natural transformation, cells enter a physiological state called competence in which they actively uptake, integrate and express genes encountered in the surrounding environment.

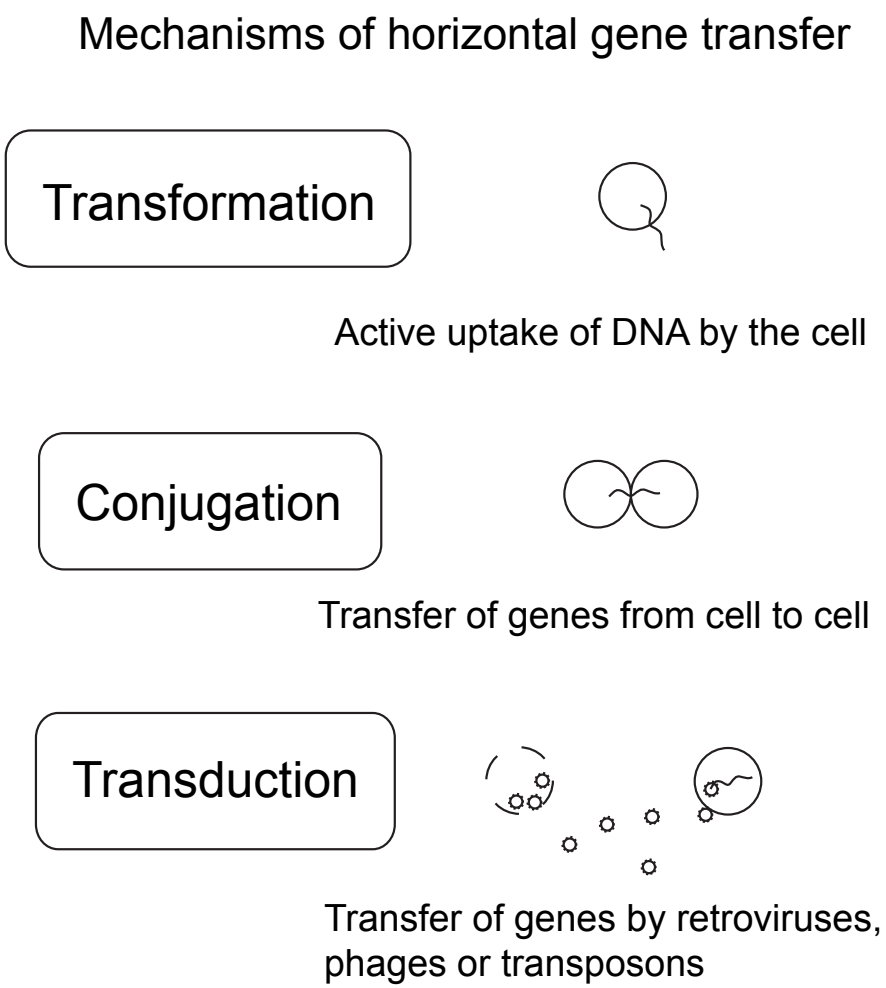

Figure 1.1: Mechanisms of horizontal gene transfer. Transformation occurs when cells take up and express DNA from the environment. Conjugation occurs between cells having specialized structures. Transduction involves the action of mobile genetic elements such as retroviruses, phages or transposable elements that carry some genes from the donor host to the recipient.

This state is usually triggered by changes in the environment. The specific trigger varies greatly depending on the species and even among strains. The ability to incorporate extracellular genes through natural transformation has been identified to occur in many representatives of both the archaea domain and in many phyla of the bacteria domains. These include Gram-positive and Gram-negative bacteria, cyanobacteria, green sulphur bacteria and many human pathogenic bacteria [21, 22]. In conjugation, genes are transferred mostly in the form of plasmids directly from one cell to another through pores or cell-to-cell junctions. For this type of horizontal gene transfer a conjugative system must be present in the donor microbe. While this type of transfer is more frequent between closely related microbial species it is not limited by relatedness. Last, transduction involves the action of mobile genetic elements such as phages, retroviruses or transposable elements. These elements are able to replicate and insert themselves into new genomes. In this process, mobile elements sometimes transport, by chance, additional genetic material from the previous host genome to the recipient genome.

1.2.3 Genome size evolution

Sizes of genomes found in archaea and bacteria are found to be smaller and vary less in genome sizes than eukaryotes. The former span only two orders of magnitude, ranging between 0.18 Mb and 13 Mb, much smaller compared to eukaryotes which span four orders of magnitude ranging between 10^2 Mb and 10^5 Mb [23, 24]. A difference found between the genomes of archaea and bacteria compared to the genomes of eukaryotes is that the bacterial/archaeal genomes tend to be much more compact. In other words, the genomes of archaea and bacteria consist almost entirely of genes with only small intergenic regions, while the genomes of eukaryotes have much larger intergenic regions. Several hypotheses have been put forward to explain this feature of archaeal and bacterial genomes. One hypothesis considers that there are strong deletion biases that purges any non-essential genes [25]. In another hypothesis it is argued that purifying selection due to large effective population sizes is sufficient to purge the genome non-essential intergenic regions which can have a negative fitness effect [26]. Evidence supporting these explanations can be found in the case of microbes that have become obligate parasites or endosymbionts. In the majority of these microbes, their genomes are seen to have undergone a dramatic reduction in genome size [27, 28, 29, 30]. In this process, many of the genes in the genome of such microbes start degenerating and are purged from the genome through gene loss. Such events have been demonstrated to occur experimentally [31]. In that study, reductions in genome size of the bacterium *Salmonella enterica* were found to occur extremely quickly in just a matter of weeks. These observations seem at odds with recent findings which show that many bacteria are robust to gene knockouts, something that would not be

expected if bacterial genomes carry only genes kept through selection. This topic will be discussed further in section 1.6.2 on the robustness of microbial organisms.

1.3 Metabolism

At any given time, hundreds of chemical reactions occur inside any living cell. These biochemical reactions allow a cell to transform the nutrients found in its environment into the small molecules needed for its growth, maintenance and function. This process is collectively referred to as the metabolism of a cell. In general, the most important of these molecules (also known as metabolites) are the units used in the building of the cell membrane (glycolipids), the transcription of genes (RNA nucleotides), the production of proteins (aminoacids), the replication of the genome (DNA nucleotides) and others necessary in some of these processes (cofactors). The study of metabolism began with the observation that microbes were responsible for the process of fermentation. Further investigation showed that this process occurred even when adding the extracts of the microbes to a sugar rich medium. By isolating and identifying the single components found in these extracts, it was observed that each component (now known to be enzymes) catalyzed a single chemical reaction in a series of reactions that used the products of the previous reaction as substrate for the next. These series of reactions, that transform one metabolite into a final metabolite through several intermediate metabolites, are known as metabolic pathways. In this manner, several pathways such as glycolysis, the Calvin cycle, and the pentose phosphate pathway were identified that serve specific and important tasks in the life of a cell. Figure 1.2 shows the biosynthesis pathway of the aminoacid serine starting from an intermediate of the glycolysis pathway. These pathways were thought to exist in many different organisms with little change, however this has recently been shown to be mostly inaccurate, except in the case of eukaryotes [32, 33].

1.3.1 Pathway evolution

The creation of a phylogenetic tree for microbes (section 1.2.1), the recent availability of completely sequenced genomes and exhaustive gene annotations for these genomes has made it possible to explore and test different hypotheses of pathway evolution. A problem that any hypothesis of pathway evolution needs to solve is the mechanism by which selection can guide the stepwise evolution of pathways. Only a complete pathway should confer a benefit to an organism, therefore incomplete pathways would be left to evolve at random under no selective pressure. One hypothesis proposed by Horowitz [34], also known as the retrogade evolution hypothesis, addressed this problem by stating that pathways evolved

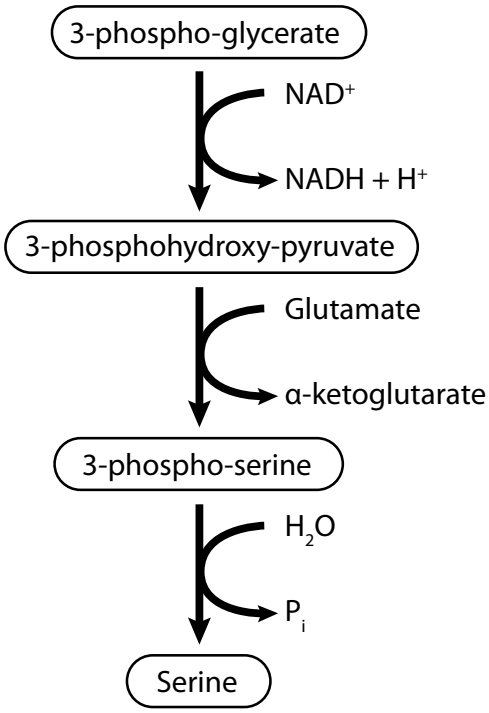

Figure 1.2: Biosynthesis pathway of the aminoacid serine starting from 3-phosphoglycerate, an intermediate in the glycolysis pathway.

through stepwise and sequential recruitment of new enzymes in the reverse order of the pathway. In other words, enzymes were recruited when they produced the final useful metabolite for the cell by transforming it from a previous intermediate metabolite A. At some point another enzyme was recruited that produced metabolite A from a previous intermediate metabolite B, and so on. This process would then lead to the evolution of a complete pathway, solving the problem of how selection would maintain the incomplete pathway. Partial pathways always produced the final beneficial metabolite from the previous intermediate metabolites and thus selection maintained the partial pathway. However, one weakness of this hypothesis is that it assumes that intermediate metabolites are readily available in the environment. This is not always the case, because some chemicals have very short lifetimes and therefore would not be taken up by a cell in the absence of specialized transport systems. Horowitz later proposed that recruited enzymes resulted from gene duplications of the enzyme coding genes which already catalyzed reactions in a pathway [35]. A more recent hypothesis proposed by Jensen [36], also known as the patchwork hypothesis, was put forward more than three decades later. Jensen addressed the weaknesses in the retrograde evolution hypothesis by arguing that enzymes are much less specific than was appreciated at the time. Recent experiments have shown that this lack of specificity does indeed exist in some enzymes [37, 38, 39]. Because of the lack of specificity, enzymes could catalyze similar reactions using different substrates with the consequence that many different pathways were latent even in organisms with limited enzyme resources. Why such latent reactions were not readily observed can be due to cell regulation which prevents unwanted reactions from occurring. However, the activity of such latent reactions could in principle be increased by increasing the levels of enzyme expression or by allowing substrate to reach high concentrations. When a product of such a latent pathway becomes beneficial, then changes to the genome that increase the production of such a metabolite will be selected. Jensen therefore proposed that new enzymes are not recruited from within the evolving pathway but from different pathways already present in the genome. The recent availability of complete genome sequences has allowed these hypotheses to be tested [40, 41]. The results show that while some limited examples of retrograde evolution can be found, the driving force of pathway evolution is the recruitment of single enzymes from different pathways [41]. This was shown by investigating whether homologous pairs of proteins are found more often within the same pathway or across different pathways in the genome of an organism. Finding a homologous gene pair in a genome indicates that the two genes diverged from a common gene after a gene duplication event. Because genes that do not evolve a novel function or confer a benefit to the organism tend to be purged from the genome, such genes probably are maintained due to the recruitment of one of the copies of the gene to a novel

function.

1.4 Genome-scale metabolic networks

The set of biochemical reactions that can occur in the metabolism of a cell can be represented as a metabolic network. In such a network, nodes represent individual chemical species (metabolites) and the links connecting the nodes represent the reactions that transform a set of metabolites (substrate) into another set (products). Because each biochemical reaction is catalyzed by an enzyme coded in the organism's genome it is possible to characterize the reactions that may occur in a cell of a microbial species by examining the set of enzymes encoded by genes in its genome. The advent of genome sequencing and the accumulation of information on enzyme function made it possible to go one step further and construct complete genome-scale metabolic models of organisms. One of the first models was of *Haemophilus Influenzae* [42], and soon after the model of *Escherichia coli* was constructed [43] with many more models assembled since.

1.5 Simulation of biochemical processes

Our understanding of an organism, and indeed of any system, is only as good as the predictive ability of the model we use. Thus, the comparison of a model's predictions and experimental results makes it possible for us to realize how much is still unknown. Another reason why it is important to develop models is that without a good model, engineering is impossible. Good models allow us to know what needs to be done such that a desired output is obtained from a system. This approach forms the basis of much biological research that has direct effects on the society of today: from environmental conservation to the many biotechnological applications. The theory that describes the dynamics of chemical reactions is known as chemical kinetics. This theory describes how the concentrations of the substrates and products involved in a reaction change over time. The changes in concentrations depend only the reaction rate, which depends itself on many variables such as the type of reaction, the concentrations of the chemical species and the free energy difference between the substrates and the products. The most common reaction type found in chemistry is the first order reaction:

$$A \xrightarrow{k} B + C \qquad \frac{d[A]}{dt} = -k[A], \frac{d[B]}{dt} = k[A], \frac{d[C]}{dt} = k[A]. \tag{1.1}$$

The reaction shown above represents a chemical species A as substrate which reacts to give two other chemical species B and C as products. The rates of change in

concentrations of the substrate $[A]$ and products $[B]$ and $[C]$ vary linearly at rate k with the concentration of the substrate.

1.5.1 Michaelis-Menten kinetics

In living organisms, the most common reaction type is not the first order reaction, but a reaction described by the Michaelis-Menten equation. A defining characteristic of these reactions is that enzymes catalyze them. Enzymes are able to accelerate the rate of reactions in many cases by a factor of more than a million. This is achieved by the formation of an enzyme-substrate complex that lowers the activation energy of the reaction. To derive the Michaelis-Menten equation we need only consider that the catalyzed reaction consists of two first order reactions: one reaction that forms the enzyme-substrate complex and another reaction in which the enzyme releases the products. The resulting equation can be further simplified by assuming that the second reaction is fast, which is usually the case for most catalyzed reactions.

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P \tag{1.2}$$

Which written in terms of rates of change of concentration becomes

$$\frac{d[ES]}{dt} = k_1[E][S] - (k_{-1} + k_2)[ES], \qquad \frac{d[P]}{dt} = k_2[ES]. \tag{1.3}$$

Assuming that the concentration of the intermediate [ES] to is in the steady state

$$\frac{d[ES]}{dt} = 0, \tag{1.4}$$

allows us to obtain the following equation for $[ES]$:

$$[ES] = [E][S]\frac{k_1}{k_{-1} + k_2}. \tag{1.5}$$

Additionally, the concentration of the enzyme $[E]$ is related to the total amount of enzyme $[E]_T$ by

$$[E] = [E]_T - [ES]. \tag{1.6}$$

At this point it is useful to introduce two constants that characterize Michaelis-Menten kinetics, the Michaelis constant K_M and the maximal rate V_{max}:

$$K_M = \frac{k_{-1} + k_2}{k_1} \qquad V_{max} = k_2[E]_T \tag{1.7}$$

Which substituted into equation 1.5 lead us to

$$[ES] = [E]_T \frac{[S]}{[S] + K_M}. \tag{1.8}$$

And finally the Michaelis-Menten equation for the reaction rate:

$$\frac{d[P]}{dt} = V_{max} \frac{[S]}{[S] + K_M}. \tag{1.9}$$

The fact that the reaction depends on the amount of enzyme available has one important consequence: at low substrate concentration the reaction is first order and its rate is directly proportional to substrate concentration; on the other hand, at high substrate concentration the reaction rate is almost constant depending little on the substrate concentration.
These effects can be seen in the plot of reaction velocity versus substrate concentration in Figure 1.3, where the reaction rate increases almost linearly at very low concentration rates, and approaches the maximal velocity V_{max} at higher substrate concentrations.
When simulating more than one enzyme catalyzed reaction such as a pathway or a larger metabolic network, the approach is to write down the set of differential equations that describe the changes in the concentrations of all the metabolites in the system and find the solution by numerical integration starting from a set of initial conditions. This approach has been used successfully in studying the reaction dynamics in pathways [44, 45]. In the case of large metabolic networks, this approach has several problems: numerical methods suffer from instabilities, and knowledge of all the kinetic rates of the reactions is required to obtain meaningful results.

1.5.2 Flux balance analysis

For genome-scale metabolic networks, the use of simplified models such as flux balance analysis [46] have been used successfully in the prediction of viability and growth rate after gene knockout [47, 48, 49]. The flux balance analysis (FBA) method considers reactions in the metabolic network to be in a steady state. In this state, the reaction rates or fluxes and the concentrations of the metabolites remain

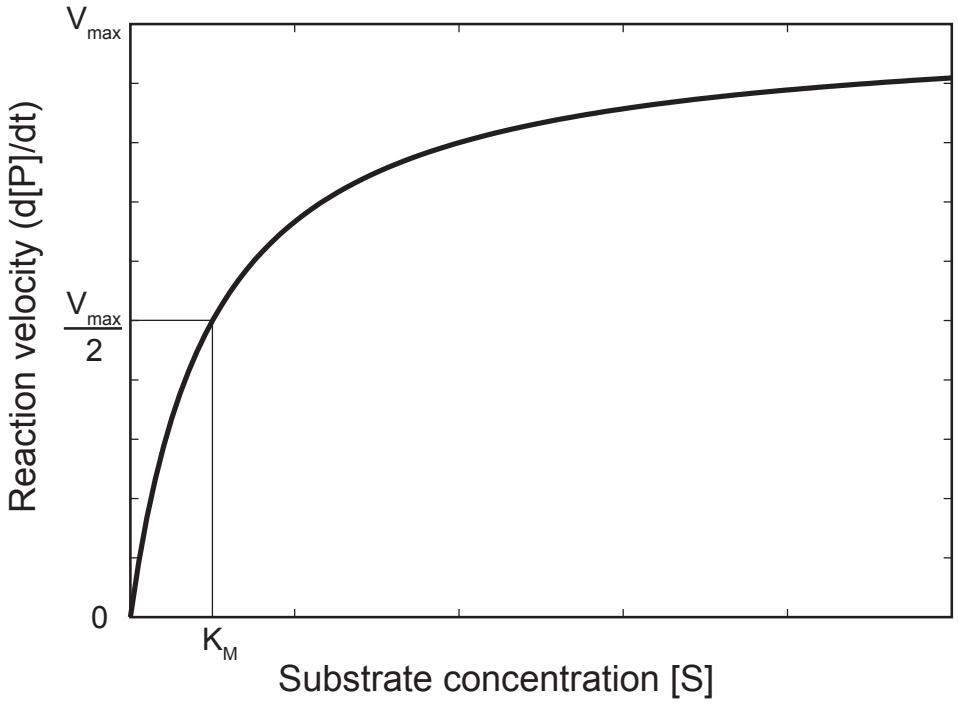

Figure 1.3: Plot of reaction velocity d[P]/dt as a function of the substrace concentration

constant in time. A consequence of this is that the production and consumption of any metabolite, by all the reactions it is involved in, must be balanced such that no net increase or decrease of the metabolite occurs. In mathematical terms, this translates to

$$\frac{d[M_i]}{dt} = \sum_{j=0}^{n} S_{ij} v_j = 0. \tag{1.10}$$

Where $[M_i]$ is the concentration of metabolite M_i, S_{ij} is the stoichiometric coefficient of metabolite i in reaction j and v_j is the reaction flux through reaction j. Another way to write this is

$$S.v = 0. \tag{1.11}$$

Where S is the stoichiometric matrix and v is the flux vector. The information needed for an FBA simulation comprises only the stoichiometric coefficients of the reactions, information that is readily available in online databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [50]. Many solutions satisfy the constraints imposed by FBA. In principle, a cell can be in any of these states. However, many of these states do not have a biological meaning. Thus, this space of solutions must be reduced to a more sensible set. This can be accomplished by measuring some of the fluxes in the organism such that other fluxes in the model are constrained. Such experiments are difficult to perform, and thus relatively little data is available on internal fluxes in organisms. Another approach that does not require experiments is by assuming that organisms maximize their cell growth or biomass production. This assumption narrows the range of solutions obtained with FBA and make predictions that are in good agreement with experiments [51]. In cases where the predictions did not match the observed growth rates, it has been shown that E. coli subject to selection pressure on rapid growth quickly evolves until it reaches the optimal growth rate predicted using FBA [52]. The adaptations are likely at the regulatory level which suggest that the regulation of metabolism can evolve very quickly to adjust an organism's metabolism to a novel environment such that it achieves its maximum growth rate. To predict the optimal growth rate a biomass reaction is needed. This reaction represents a cell's consumption of biomass precursors (cofactors, glycolipids, nucleotides and amino acids). The exact coefficients of this reaction are estimated based on cellular biomass composition [53, 54]. Finally, one needs to specify the maximum uptake rate of nutrients, which depends on the nutrients available in the environment. With this information one can then apply linear programming [55] to find the

solution to the maximization problem corresponding to

$$Max(v_{biomass}), \tag{1.12}$$

given certain bounds on the consumption of environmental available nutrients

$$0 \leq v_{ext_k} \leq E_k, \tag{1.13}$$

where v_{ext_k} are the flux entries in v that correspond to the exchange of metabolites with the environment and E_k is the maximum rate at which those metabolites can be consumed. FBA has also been successful in predicting other aspects of metabolism such as metabolite secretion, regulation of metabolism and reaction essentiality [42, 47, 56, 51, 57, 48, 58, 59]. I have used this approach to explore two questions regarding the evolution of metabolic networks: 1) the role of constant phenotype networks in metabolic network evolution and 2) whether the robustness found in *E. coli* is typical of random viable metabolic networks.

1.6 Genotype-phenotype maps

The discovery that DNA is the molecule that contains all the information necessary to produce a new organism established the first link between the genotype (the information contained in the DNA molecule) and the phenotype (the characteristics of the organism produced). The development of genome sequencing methods gave us access to the information encoded in the DNA and allowed us to investigate how this information is used when new organisms are developing. Recent experimental techniques have enabled us to manipulate the information contained in DNA, thereby changing the phenotype of the organism. Any function that associates genotypes with a corresponding phenotype is called a genotype-phenotype map [60]. While it is possible to manipulate experimentally the genetic information of an organism, it is impossible to do this at a sufficiently large scale such that general properties of these genotype-phenotype maps can be studied. Some of these properties have important implications for the ability of organisms to evolve new phenotypes, as will be discussed bellow. This experimental limitation has been overcome using computational methods, because it is currently possible to predict the phenotypes given a genotype for some classes of biological components. Examples include protein structure [61], RNA secondary structure [62] and regulatory networks [63]. I present in Chapters 2 and 3 a study of the genotype-phenotype maps of metabolic networks and their ability to produce biomass precursors from different environments. In these studies, I show that some properties of the genotype-phenotype maps of metabolic networks are found to be common with other maps, while other properties are different.

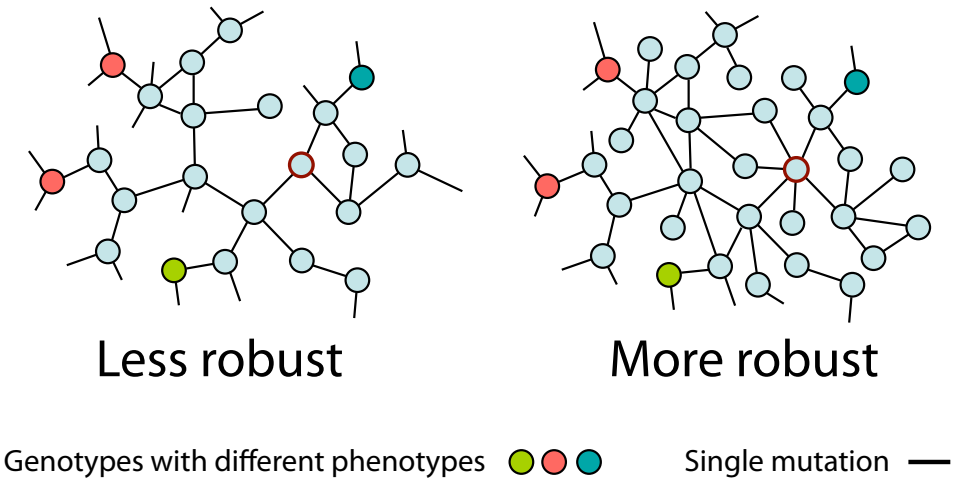

Figure 1.4: Illustration of a constant-phenotype genotype network. Nodes in the network represent genotypes and links indicate that the connected genotypes differ by a single mutation. Phenotypes are shown as the colors inside the circles representing genotypes. In many biological systems, genotypes can evolve through single mutations while keeping their phenotype constant. This is illustrated here through the interconnected blue nodes. When genotypes are more robust, more mutations from a single genotype lead to other genotypes with the same phenotype.

1.6.1 Constant-phenotype genotype networks

One property found common across the genotype-phenotype maps of biological systems studied so far is the existence of networks of genotypes which all share the same phenotype and can be traversed through single mutations. These constant-phenotype genotype networks illustrated in Figure 1.4 are often referred to as neutral networks or simply genotype networks. Genotype-phenotype maps of biological components in which such genotype networks are found are: RNA [62], proteins [64] and regulatory networks [63]. Experimental verification of the existence of these networks is still limited in scope but has been shown in protein evolution experiments [65, 66], and supported by phylogenetic studies of protein function [67]. The existence and characteristics of these networks have been shown to play a fundamental role in the ability of organisms to encounter novel phenotypes. This point will be discussed further in section 1.6.3 on innovation. In chapters 2 and 3, I explore the existence of these networks and their effect on the ability to acquire innovations found in the case of metabolic networks.

1.6.2 Robustness

The second point investigated in this thesis in chapters 2 and 3, regards the origin of biological robustness of metabolic networks. Generally, robustness refers to the ability of a system to maintain its function despite perturbations. Biological systems that maintain their function in spite of perturbations have an advantage over biological systems that lose their functionality in the same case. Two types of robustness can be observed in biological systems. One is genetic robustness, which is the constancy of phenotype in the presence of heritable perturbation such as mutations of the genomic information. Another is environmental robustness, which is the buffering of perturbations that are nonheritable in origin. These can be any external environmental factors such as temperature or salinity. The specific mechanisms responsible for an organism's robustness depend on the type of perturbation and the biological system in question [68, 69]. Regulatory networks are involved in many important functions. Examples include the response to, and detection of, signals in chemotaxis, the control of circadian clocks, as well as the regulation of cell cycle and organism development. The explanation for the robustness of regulatory networks to noise can be attributed to the existence of positive and negative feedback in the networks. In systems control engineering, it is known that feedback mechanisms increase the signal quality. One source of genetic robustness is the existence of alternative or fail-safe mechanisms that provide the same function [68]. They ensure that if one mechanism fails, the others will maintain the function as a whole. This type of robustness is also known as redundancy. One problem that has drawn much interest in the study of robustness

is the evolutionary explanation for its origin. There are three main views on how robustness can arise as a product of evolution [69]. First, robustness can evolve because a more robust genotype produces more viable offspring thereby having a positive effect on fitness. Second, it can evolve intrinsically because the selection acting on a trait that increases an organism's fitness is correlated with robustness. And third, robustness may be correlated with environmental robustness which selection would act to maximize.

Constant-phenotype genotype networks in which some genotypes are more robust than others have been shown to increase the average evolved robustness in an organism [70]. This occurs only when the mutation rate is high enough and when there is a significant cost in fitness incurred from the production of inviable offspring. This may be the case in RNA viruses [71, 72], but it is not the case in prokaryotic genomes where mutation rates are low. In chapters 2 and 3, I investigate the genetic robustness of randomly evolved metabolic networks. In chapter 2, the robustness of randomly evolved metabolic networks is compared to the robustness found in the metabolic network of E. coli. Here I focus on robustness to gene knockout or loss of function point mutations. It has been shown experimentally that microbes such as E. coli or S. Cerevisiae are robust to gene deletions [47, 48, 49]. In metabolic networks, two types of mechanisms that provide robustness [73] can be found: 1) redundancy through gene duplicates or enzymes that catalyze the same reaction and 2) through distributed robustness, in which different pathways are able to maintain the function of a pathway that has lost its functionality through mutation.

1.6.3 Innovation

The ability of an organism to find novel phenotypes may be heritable. This is intuitive if we consider genotype networks. In such a network, the genotypes that a specific genotype can reach through a single mutation consist of its first neighbors. This neighborhood and the phenotypes associated with it do not change and are therefore a property of the genotype. The existence of networks of interconnected genotypes which share the same phenotype drastically increases the ability of an organism to encounter novel phenotypes. This implies that an organism can reach a very different genotype through single mutations while maintaining its phenotype, potentially allowing it to encounter novel phenotypes. One important requirement for genotype networks to facilitate innovation is that the phenotypes found in the neighborhoods of two genotypes vary quickly with the number of mutations between genotypes. This has been shown to be true in the biological systems studied so far, and I show that this also occurs in the case of metabolic networks in chapters 2 and 3.

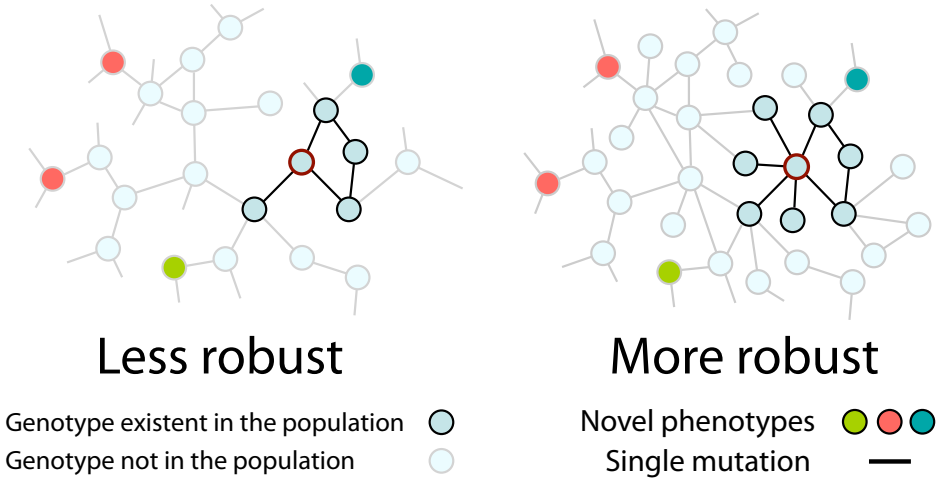

Figure 1.5: Illustration of genotypes in evolving population in two different genotype networks. Nodes in the network represent genotypes and links indicate that the connected genotypes differ by a single mutation. Bright colored genotypes represent genotypes existing in the population while faded genotypes indicate the genotype is not present in the population. In less robust genotype networks populations become less diverse and may therefore encounter smaller numbers of novel phenotypes when compared to more robust genotype networks.

1.6.4 Robustness and innovation

High robustness is in principle incompatible with the ability to innovate [74]. This observation is straightforward if we consider that robustness reduces the effect of perturbations on a system. This reduction should then affect equally both deleterious phenotypes as well as novel, potentially beneficial, phenotypes. Since both robustness and the ability to innovate can be seen to exist in biological organisms, this implies that two independent and opposing evolutionary forces must drive robustness and the ability to innovate such that a balance is reached between them [75]. One alternative explanation has been described in the case of RNA molecules [74]. There it was shown that less phenotype diversity was accessible through single mutations in the neighborhood of individual RNA molecules with high robustness. However, when one considered the phenotype diversity accessible to a whole population of evolving sequences with conserved phenotypes this incompatibility was resolved. In this case, robustness was seen to correlate positively with the accessible phenotypes of all sequences of the population. The explanation for this result can be understood with the help of Figure 1.5 and was proposed by Wagner [76, 74]. The genotypes found in an evolving population can be seen in the case of two different hypothetical genotype networks. In the least robust genotype network, individuals tend to cluster around fewer genotypes due to the more frequent production of inviable offspring. On the other hand in more robust genotype networks, the population becomes more diverse and is capable of exploring more genotypes and therefore encounters a greater number of novel phenotypes. To confirm if this is indeed a general characteristic found in biological systems, I examine here if this can also be seen in the case of metabolic networks. If robustness and the ability to innovate are generally found to correlate this implies that only one evolutionary force driving both robustness and the ability to innovate is needed to explain the observed robustness in organisms. However, in chapter 3 I show that this is not the result found in the case of the robustness of metabolic networks.

1.7 Evolution of terminal and reversible cell differentiation

Several major transitions mark the evolution of life as we know it. During these transitions, fundamental changes in the level at which natural selection acted on lead to the evolution of organisms with greater complexity [77, 78, 79]. One of these transitions was the origin of multicellularity. The fact that this transition occurred several times in independent lineages suggests that it is an easy transition [80]. Many evolutionary forces may have driven this transition. Some examples are selection for larger size, group metabolic effects [81] or benefits from the ability to differentiate once multicellularity was established.

1.7.1 Terminal and reversible differentiation

In my thesis, I concentrate in chapter 4 on the evolution of terminal and reversible differentiation, a problem that has received little attention so far. Differentiated multicellular organisms are composed of many cell types. In most of these organisms, a precise location of the different cell types is a necessary condition for their proper function in the organism. Consequently, the development of offspring in differentiated organisms is more complex than in single celled organisms. Almost all multicellular organisms are able to develop from a single cell, however in most organisms not all cell types have the ability to originate new organisms. Instead, usually organisms develop from a single cell type. In this type of reproduction it is possible to make a distinction between the cells that pass the genetic information to the next generation (germline) and the ones whose genetic information is lost when the organism dies (somatic). In another form of reproduction, known as vegetative reproduction or reproduction by fragmentation, different cell types or small groups of cells are able to develop new organisms. In this form of reproduction, there is no distinction between germline and somatic cells. Despite the potential benefit of this form of reproduction, it is not found in organisms as often as would be expected. It occurs in very few animals [82, 83, 84] and while it occurs in many plants, in most of them it is observed only in special conditions [82]. This observation is unexpected because in almost every organism, all different cell types contain the complete genomic information, so no a priori reason exists that explains why organisms would not evolve this ability. An organism's ability to reproduce through fragmentation is tightly connected with the ability of the composing cell types to differentiate into other cell types. Cell types that are unable to differentiate into other cell types are referred to as terminally differentiated cells. If a cell type cannot differentiate into all other cell types, it cannot originate a new organism and thus such organism cannot reproduce through fragments consisting

only of terminally differentiated cells. In another case, cell types have the ability to differentiate from one type to the other and are therefored referred to as reversibly differentiated. An organism composed of reversibly differentiated cells can potentially reproduce through fragmentation consisting of any type of cells. One question that will be further examined in chapter 4 regards the conditions that determine which cell type becomes the germline. In that chapter, it will be shown that the relative growth rate is the single most important factor determining the cell type that becomes the germline. Specifically, the cell type with the fastest growth rate evolves to acquire that role.

1.7.2 Photosynthesis and nitrogen fixation

The study presented in chapter 4 investigates the conditions that affect the evolution of these two types of cell differentiation. There, I develop a model based on the particular setting of two interacting cell types: cells that specialize in nitrogen fixation, and cells that perform photosynthesis. Photosynthesis and nitrogen fixation are two incompatible processes because oxygen irreversibly disables nitrogenase [85]. The solution to this incompatibility requires their separation. Many strategies have evolved in different organisms to accomplish this separation. These range from temporal separation such as the one observed in unicellular circadian cyanobacteria [86], to spatial separation as observed in multicellular differentiated cyanobacteria [87, 88, 89]. Symbiotic relationships such as the one found between most plants and cyanobacteria or rhizobia are a form of spatial separation [90].

1.7.3 Multicellular cyanobacteria

Differentiated multicellular cyanobacteria are well known organisms composed of cells that specialize in photosynthesis and others that specialize in nitrogen fixation. Cyanobacteria in general, have evolved many different morphologies and are found to exist as single cells, as multicellular filaments of undifferentiated cells, as well as differentiated multicellular linear or branched filaments [91]. In the case of reversibly differentiated cyanobacteria, only the *Trichodesmium* cyanobacterium has not been observed to have terminal differentiation. While no distinction can be made morphologically between the cells of these cyanobacteria, experiments show that cells are differentiated in their function due to differences in protein expression [89, 92]. In contrast, many terminally differentiating cyanobacteria are known. Two examples are *Anabaena* and *Nostoc*. In these types of cyanobacteria two cell types can be distinguished visually on the microscope: the vegetative cell (germline) and the heterocyst cell (somatic). The vegetative cell is photosynthetic, reproduces through division and differentiates into a heterocyst cell when fixed nitrogen is not available in the environment [93]. The heterocyst cell is larger than

the vegetative cell, has a thicker cell wall composed of three layers, and is the cell that performs nitrogen fixation. When fixed nitrogen is lacking in the environment, some vegetative cells differentiate into heterocyst cells. In this manner, vegetative cells obtain fixed nitrogen from heterocyst cells and heterocyst cells obtain fixed carbon from the vegetative cells.
In chapter 4 I show that while the topology of cell interactions and differentiation costs play a role in the type of differentiation evolved, differential cell growth rates are the main factor determining the evolution of terminal or reversible differentiation.

2 Evolutionary plasticity and innovations in complex metabolic reaction networks

João F. Matias Rodrigues and Andreas Wagner
[*PLoS Computational Biology*, **2009**, 5(12) e1000613] (doi:10.1371/journal.pcbi.1000613)

2.1 Abstract

Genome-scale metabolic networks are highly robust to the elimination of enzyme-coding genes. Their structure can evolve rapidly through mutations that eliminate such genes, and through horizontal gene transfer that adds new enzyme-coding genes. Using flux balance analysis we study a vast space of metabolic network genotypes and their relationship to metabolic phenotypes, the ability to sustain life in an environment defined by an available spectrum of carbon sources. Two such networks typically differ in most of their reactions and have few essential reactions in common. Our observations suggest that the robustness of the *Escherichia coli* metabolic network to mutations is typical of networks with the same phenotype. We also demonstrate that networks with the same phenotype form large sets that can be traversed through single mutations, and that single mutations of different genotypes with the same phenotype can yield very different novel phenotypes. This means that the evolutionary plasticity and robustness of metabolic networks facilitates the evolution of new metabolic abilities. Our approach has broad implications for the evolution of metabolic networks, for our understanding of mutational robustness, for the design of antimetabolic drugs, and for metabolic engineering.

2.2 Summary

Understanding the fundamental processes that shape the evolution of bacterial organisms is of general interest to biology and may have important applications in medicine. We address the questions of how bacterial organisms acquire innovations, including drug resistance, allowing them to survive in new environments. We simulate the evolution of the metabolic network, the network of reactions that can occur inside a living organism. The metabolic network of an organism depends on the genes contained in its genome and can change by gaining genes from other organisms through horizontal gene transfer or loss of gene activity through mutations. Our observations suggest that the robustness to gene loss in *Escherichia coli* is typical of random viable metabolic networks of the same size. We also find that metabolic networks can change significantly without causing the loss of an organism's ability to survive in a given environment. This property allows organisms to explore a wide range of novel metabolic abilities, and is the source of their ability to innovate. Finally we present a method to find reactions that are essential across all organisms. Drugs targeting such a reaction may avoid drug resistance mutations that bypass the reaction.

2.3 Introduction

Organisms, especially microbes, thrive on organic nutrients with bewildering diversity: the vast majority of organic molecule can mean "food" for some species. From a microbe's perspective, acquiring the ability to survive on a new carbon source can make the difference between life and death; such an acquisition can thus be an important evolutionary innovation. We here study the properties of metabolic systems that facilitate such innovations. The evolution of biological macromolecules has received serious attention for decades [94]. The same is not true for biological systems on higher levels of organization, such as regulatory and large complex metabolic networks. One reason is a comparative paucity of data for such networks. Another reason is the inherent difficulty in characterizing both network genotypes and network phenotypes. Recent work on genome-scale metabolic networks reduces these limitations. First, metabolic genotypes have recently been characterized for several model organisms [42, 95, 96]. Second, databases of metabolic reactions inform us about a broad spectrum of chemical reactions catalyzed by enzymes in living things. Third, flux balance analysis [97] allows us to compute metabolic phenotypes from metabolic genotypes (Figure 2.1). Taken together, these developments allow us to study the evolution of metabolic networks in greater depth. The functions and phenotypes of biological macromolecules are robust to genetic change. Such robustness has important implications for the evolutionary plasticity of molecules, the ability of molecules to evolve new properties. Through mutations that do not affect a molecule's function, vast regions of phenotype space can be explored, regions in which molecules with novel phenotypes can lie [94, 98]. Does the same hold for genome-scale biological networks? Can biological networks with similar phenotypes have a vast number of interconnected and different genotypes, thus being both highly robust and having large evolutionary plasticity? These questions currently have few answers. We study the evolution of genome-scale metabolic networks to provide such answers. For our purpose, a metabolic genotype is a set of chemical reactions catalyzed by gene-encoded enzymes that take place in an organism. Any one organism's metabolic network exists in a much larger space of metabolic genotypes. This space is defined by the biochemical reactions known to be realized in living cells. Any one organism's genotype can be thought of as a point in this space, where some biochemical reactions occur and others are absent. Genotypes can thus be represented as binary strings whose entries indicate presence ('1') or absence ('0') of reactions (Figure 2.1) in an organism. We define the phenotype of such a network as its ability to sustain life in a given environment or set of environments. This means that the network must be able to produce all biochemical precursors (amino acids, nucleotides etc.) that are necessary to allow a free-living heterotrophic organism such as Escherichia coli to grow from environmental resources. We here

consider 101 minimal environments that only differ in their carbon source. Specifically, these environments provide only a terminal electron acceptor (O_2), a source of nitrogen (NH_3), sulfate (SO_4), phosphate (PO_4), and one out of 101 sources C of carbon (see section 2.7 for a complete list of all carbon sources used). We can represent a metabolic phenotype as a binary string, whose i-th entry is equal to one (Figure 2.1), if a network is able to sustain life when C_i is the only available carbon source. A network able to sustain life in complex environments with multiple carbon sources has phenotypes in which many of these entries are equal to one. Metabolic phenotypes, as defined here, can be computed from metabolic genotypes using flux balance analysis. Flux balance analysis is a computational tool that relies both on stoichiometric information about chemical reactions occurring in a cell, as well as on an objective function such as the production of biomass precursors. For a given nutritional environment, it computes allowable rates at which individual reactions proceed in a metabolic steady state, and these rates in turn determine whether all necessary biochemical precursors can be produced. Its qualitative predictions growth or no growth are in good agreement with experimental data for well-studied model systems [49, 56]. We here study the evolution of metabolic networks in the space of the genotypes just defined. Genotypes can change through the elimination of chemical reactions caused by loss of function mutations in enzyme-coding genes. Many such mutations do not abolish a network's ability to sustain life [96, 49, 56, 99, 100, 101, 102, 103, 104, 105, 106, 107]. Genotypes can also change through addition of chemical reactions, which occurs at appreciable rates in prokaryotes through horizontal gene transfer [99, 17]. This motivates our choice of a prokaryotic network that of E. coli as the departure point of our work [95]. Two further reasons compelled us to choose specifically the *E. coli* network. First, it is perhaps the most prominent and well-studied example of a metabolic network in a free-living organism. Second, more effort has been devoted to studying its robustness than for other networks [49, 56, 48, 106, 108, 109, 110, 111, 112, 113]. For these reasons we also wanted to compare properties of the E. coli metabolic network with those of the sampled networks that our approach generates. Mutations and horizontal transfer can sometimes affect more than one enzyme-coding gene (reaction), but we focus here on the individual reaction as the elementary unit of change. Each such change transforms a network into one of its immediate neighbors differing from it by one reaction. We refer to all of a network's neighbors as a network's neighborhood. Methodologically, our approach bears resemblance to that of an earlier study [106] which asked how minimal genomes evolve from the E. coli genome through metabolic gene loss. However our method is new in that we do not limit ourselves only to the elimination of chemical reactions but 1) we allow for the addition of metabolic reactions, which allows us to explore a vast genotype space, 2) our analysis is not limited to *E. coli*, and 3) we also explore a very large number

of different environments. In this context, we ask several fundamental questions about the organization of genotype space, and about the ability of metabolic networks to find evolutionary innovations in this genotype space. How different can the organization of two metabolic networks be while still preserving similar phenotypes? How many mutational steps are needed to get from a network with a given phenotype to one with a very different phenotype? How different are the new phenotypes that a network encounters in its immediate neighborhood during evolution? The answers to these questions can not only elucidate why metabolic networks are robust to mutations [96, 49, 56, 102, 114, 115, 116, 117, 118]. Even more importantly, they also tell us how metabolic innovations can arise through a metabolic network's exploration of a vast space of possible genotypes.

2.4 Results

2.4.1 Networks supporting life in one environment can have very different essential reactions

We begin our analysis with a simple phenotype, a metabolic network's ability to produce all biochemical precursors from a single carbon source, glucose, in an aerobic minimal medium (see section 2.7 for a list of all environmental metabolites). The *E. coli* metabolic network [95], excluding 205 transport reactions, catalyzes 726 out of the universe of 5870 reactions we consider (see section 2.7 for details on reaction compilations). Its immediate neighborhood in genotype space consists of the 5870 networks that differ from the *E. coli* network by one (added or eliminated) reaction. Addition of a reaction to a network would not impair its ability to grow on glucose, but elimination of a reaction might. Out of the 726 *E. coli* reactions, 210 reactions are essential and cannot be removed without abolishing growth on glucose minimal medium. Thus, only 3.6% (210/5870) of the entire neighborhood, and only 29% (210/726) of those neighbors with one deleted reaction, are not able to sustain life on glucose minimal medium. Are metabolic networks that are very different from the *E. coli* network, but that can also sustain life on glucose similarly robust? To address this question, we analyzed 1000 such networks (Figure 2.1). These networks were the end points of 1000 long random walks of 10^4 mutational steps each through genotype space that started from the *E.coli* network. Figure 2.6 shows the evolution of genotype distance and network size in one such random walk. Each step consisted of the random addition or deletion of one chemical reaction and was required to preserve the ability to sustain life on glucose minimal medium. For brevity, we will call the end-point of such a random walk a random viable metabolic network with a given phenotype. We emphasize that the number of reactions in the random viable metabolic networks is similar to that of the E.

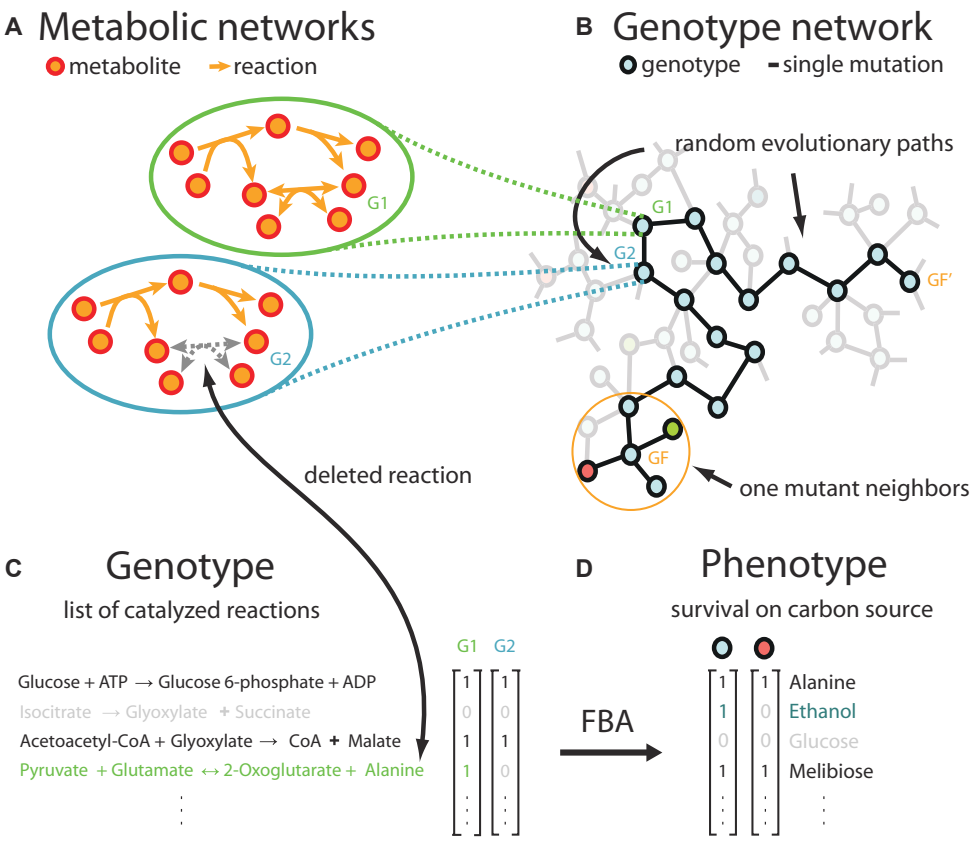

Figure 2.1: Exploration of a vast genotype space of metabolic networks. A genotype can be represented in different ways: (A) as a metabolic network, (B) as a node in a genotype network, or (C) as a binary vector listing the reactions catalyzed. Genotypes on the genotype network (B) that are connected differ by only one mutation. The color of the genotype circles indicates their metabolic phenotype. Metabolic phenotypes are computed using FBA applied to 101 environments with different carbon sources. They can be represented as a binary vector listing the environments a genotype is viable in (D). Random evolutionary walks can be seen as paths on a genotype network. Two independent random walks are shown with the same starting genotype (G_1) and two final genotypes (G_F and $G_{F'}$), passing through intermediate genotypes (i.e.: G_2) that differ by one mutation. Mutations are chosen at random. They can be additions or deletions of individual reactions from the corresponding metabolic network but they must not change the phenotype. The neighborhood of each genotype can be analyzed by characterizing the phenotype of the one mutant neighbor genotypes (approximately 5800 neighbors per genotype). The number of genotypes in the genotype space is 2^{5800}. Each genotype is able to catalyze approximately 1000 out of 5800 possible reactions.

coli metabolic network (see section 2.7 for algorithmic details). We examined the neighborhood of each of these 1000 random viable networks to identify essential reactions in them. Figure 2.2a shows the distribution of the number of essential reactions. It varies across a narrow range between a minimum of 213 (26.4%) and a maximum of 257 (32.4%) reactions. The robustness of the *E. coli* network lies in the bulk of this distribution, and is thus not atypical. This suggests that for a typical metabolic network with a given phenotype, many different mutational changes leave the network's ability to sustain life in a given environment unchanged. How different are the networks that can sustain life in this simple environment? We addressed this question in two complementary ways. First, we asked how many essential reactions differ between each network pair drawn from the 1000 random viable networks we had generated previously. Specifically, we represented the set of all essential reactions by a binary vector. For each of the 1000 random viable networks, this vector contained a 1' for a reaction that was essential in the respective network, and a 0' for a reaction that was nonessential. We calculated the normalized Hamming distance between these vectors for each pair, which is the fraction of entries at which these vectors have different values. This distance ranges from zero if a network pair has completely identical essential reactions to one if a network pair has no essential reactions in common. Figure 2.2b shows the distribution of the fraction of essential reaction that two networks have in common. On average, 32.9% of essential reactions are different in two random viable networks with the same phenotype. If we exclude reactions from this analysis that are essential in all 1000 networks, then 74% of essential reactions differ among networks. We next ranked all reactions according to the number of networks (among 1000) in which they were essential. Reactions essential in all 1000 networks received the lowest rank, and reactions that were essential in successively fewer networks received increasingly larger ranks. This ranking indirectly estimates the abundance of alternative pathways around any given reaction in a random viable metabolic network. If there are many alternative pathways, then the reaction will rarely appear as essential; if there are no alternate pathways, the reaction will appear as essential in all metabolic networks. The majority (4550) of reactions were never essential. Among the 1420 reactions that were essential in at least one network, only a small minority of 7.3% (103) reactions were essential in all networks. As an example, Figure 2.3 shows a measure of the reaction rank for a small subset of reactions, the key reactions in central energy metabolism (glycolysis, pentose phosphate shunt, citric acid cycle) color-coded according to whether they are rarely (blue) or frequently (red) essential. All of the 26 reactions were essential in more than one percent of the 1000 random viable networks. Around 46 percent of the reactions (12/26) were essential in more than 10 percent of the networks. Merely three reactions were essential in almost all of the networks. They come from gly-

colysis (glucose 6-phosphate isomerase), the citric acid cycle (aconitase), and from the pentose phosphate pathway (ribulose 5-phosphate 3-epimerase,). Note that two reactions that belong to the same (apparently unbranched) pathway of Figure 2.3 may show different essentiality. This can be understood by considering that for each reaction there may be a different number of alternative pathways (whose reactions are not shown in the figure) but that can compensate for the loss of the reaction. To validate our analysis of reaction essentiality with empirical data, we tested the following prediction: If a reaction is frequently essential in our random viable metabolic networks, then its enzyme-coding genes should also occur in a large number of different genomes. This is indeed the case, as we show in Figure 2.9. The figure demonstrates that the frequency of a reaction as essential and the number of prokaryotic genomes carrying an enzyme-coding gene that catalyzes this reaction are positively correlated (Pearson's $r = 0.45$ and $p = 2.2 \times 10^{-16}$). For this analysis we used the information available in the KEGG database [50, 119]. Taken together, these observations show that networks with the same phenotype are highly plastic in their organization. Many essential reactions typically differ between pairs of such networks. This holds even for reactions in the most central parts of metabolism.

2.4.2 Networks supporting life in one environment can have very different genotypes

In a second effort to characterize the plasticity of network organization, we asked how distant from the *E. coli* network a network can maximally be and still preserve the ability to sustain life on a glucose-minimal medium. To do so, we generated 1000 networks from the E.coli network through a random walk similar to that described above, but where we forced each step of the random walk to increase the distance to the *E.coli* network. Figure 2.4a shows that more than three quarters of genotype space can be traversed without destroying the metabolic phenotype. An environment in which metabolic networks have to synthesize every single biochemical precursor is demanding. Thus, our observations might depend strongly on the nature of this environment. However, this is not the case. We also examined a rich medium in which 36 biochemical precursors are provided for the cell (see section 2.7 for details). In such a medium, 15.9% of reactions are essential on average (13.5% fewer than in minimal medium) (Figure 2.2a); the percentage of essential reactions that differ among two networks is very similar (33.8%; Figure 2.2b); the number of reactions that are essential in at least one environment is smaller (1304 vs. 1420); a smaller percentage (5.1%; 67 of 1304) of reactions are essential in all networks (Figure 2.2c; Table S1); and the maximal distance of networks to the *E. coli* network is on average 83.9%, even larger than in minimal

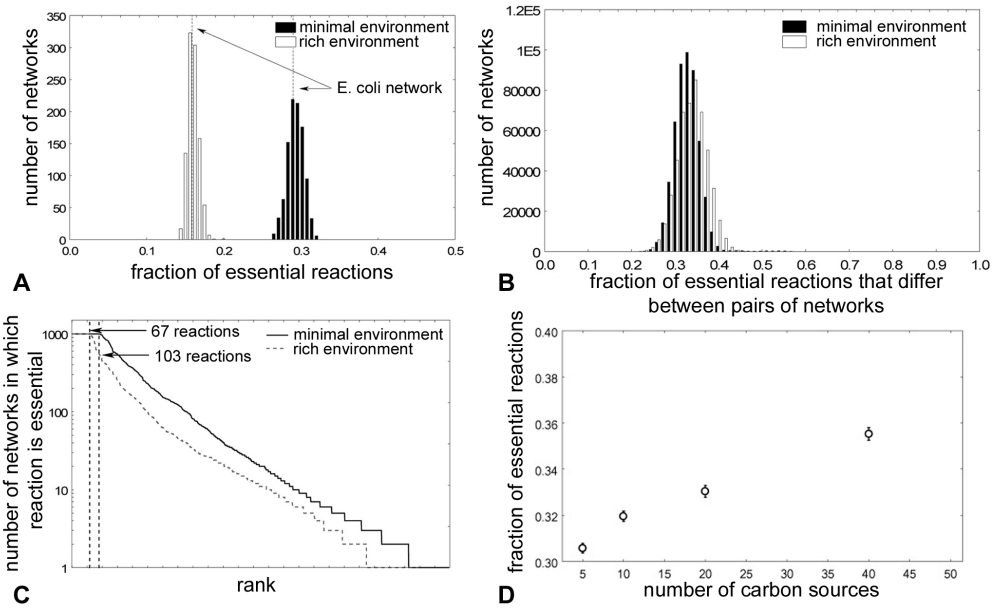

Figure 2.2: Essential reactions differ dramatically between metabolic networks with the same metabolic abilities. (A) Distribution of the fraction of essential reactions in 1000 random networks viable in minimal or rich glucose containing medium. (B) Distribution of the fraction of essential reactions shared among pairs of these 1000 random networks. (C) Rank plot of reaction essentiality. Reactions essential in all of the 1000 random viable networks are given the lowest rank of one. (D) The average fraction of essential reactions (vertical axis) as a function of the number of carbon sources a network can sustain life in (horizontal axis). Each point is an average of 100 networks (whiskers: 95% confidence interval).

medium (Figure 2.4a). Thus, evolution in a rich versus a minimal environments does not change our results dramatically. It is instructive to examine the reactions essential in all networks more closely. They are significantly enriched in reactions involved in tyrosine biosynthesis (P=0.01), cell wall biosynthesis ($P = 1.0 \times 10^{-10}$), and membrane biogenesis ($P = 2.8 \times 10^{-6}$). Taken together, the following picture emerges from these observations. Networks that have the ability to sustain life on a particular carbon source have many neighbors in genotype space with the same ability. By mutationally stepping from neighbor to neighbor (through addition and deletion of chemical reactions) network organization can change fundamentally without losing this ability. Two networks with this ability can contain very different sets of reactions, and very different essential reactions. Because networks with the ability to sustain life in a given environment are connected through their neighbors in genotype space (see section 2.7 for details) this means that large fractions of genotype space can be traversed on evolutionary time scales without affecting any one metabolic ability.

2.4.3 Metabolic networks with complex carbon phenotypes can also have very different organizations

We next turn to more complex phenotypes, namely the ability for a network to sustain life if any one of multiple carbon sources is provided in an otherwise minimal environment. We here focus on the 101 potential carbon sources annotated to have associated transport reactions in *E. coli*. Because the requirement to sustain life on an increasing number of carbon sources may increasingly constrain network architecture, our observations from above may not hold for such complex phenotypes. Figure 2.4b, however, shows that this is not the case. The figure examines the maximal genotype distance from the *E. coli* network achievable for networks with the same phenotype, as a function of the phenotype's complexity, that is, the number of carbon sources a network can sustain life on. This maximal distance declines by less than 10% for networks that can sustain life on 60 carbon sources. Thus, even if a network can sustain life in many different carbon-containing environments, its architecture is not highly constrained. The fraction of reactions that are essential does not change dramatically either (Figure 2.2d). Specifically, it increases modestly from a mean of 0.3 (Figure 2.2b) to 0.4 (Figure 2.2d) for networks that can sustain life on 5 and 60 different carbon sources, respectively. In this analysis, we used a very conservative definition of essentiality. For example, for networks able to sustain life on 60 different carbon sources, we call a reaction essential if it is required in at least one of the 60 minimal environments distinguished by these carbon sources. If we define reaction essentiality less conservatively, then the fraction of essential reactions actually decreases with an increasing number of

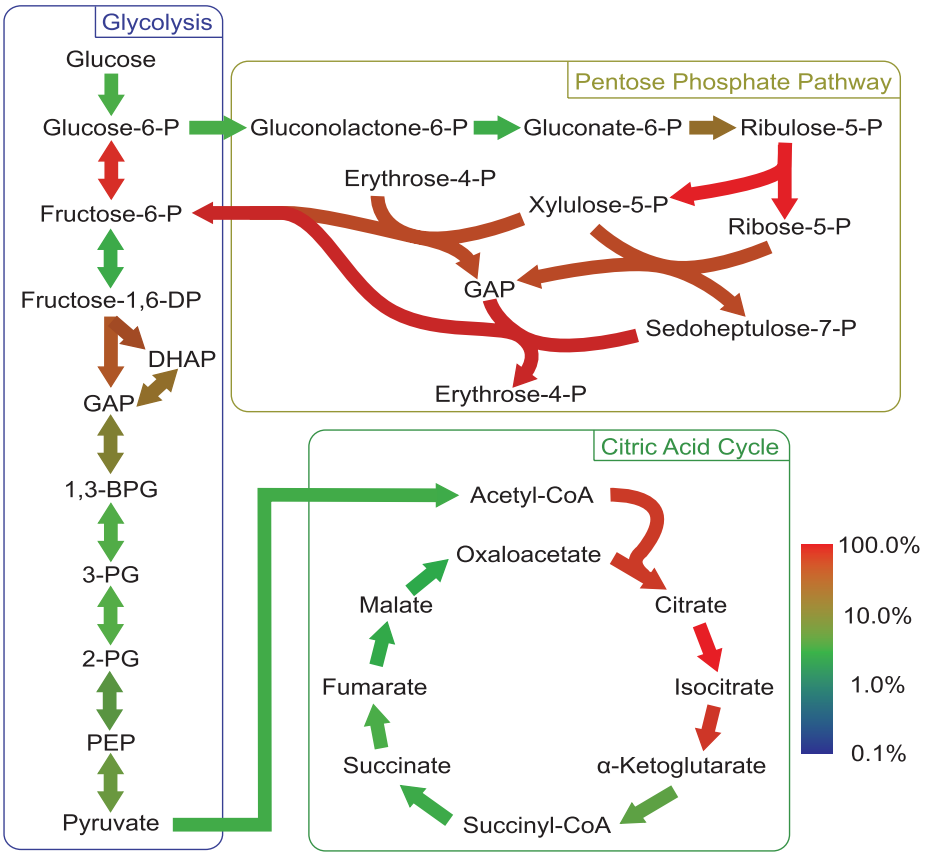

Figure 2.3: Reaction essentiality in central metabolism. Color-coded map of reactions in central energy metabolism that appear rarely (blue) or frequently (red) as essential in 1000 random viable metabolic networks. The color is in logarithmic scale indicating that most reactions even in this most central part of metabolism are essential only in a small fraction of networks with a given metabolic phenotype.

carbon sources (Figure 2.7).

2.4.4 Networks with different phenotypes can be found close together in genotype space

We next studied several properties of metabolic networks that relate to their ability to evolve new phenotypes. The first such property regards the minimum genotype distance of two metabolic networks with arbitrary, different phenotypes. If this distance is typically large, then it would be very difficult to reach any one phenotype from a network with a different phenotype through a modest number of genetic changes. To determine this distance, we first created a pair (G_1, G_2) of metabolic network genotypes with randomly chosen different phenotypes, as described in the section 2.7. We then carried out a random walk that started from G_1 and that approached G_2 in genotype space, while leaving G_1's phenotype unchanged. When this random walk had reached a point where the genotype distance to G_2 could no longer be reduced, we stopped and recorded the minimal distance thus obtained. We repeated this procedure for 1000 metabolic network pairs with different phenotypes. Figure 2.4c shows a histogram of this minimal distance for networks that are required to sustain life on at least one carbon source. It is evident from the Figure that this distance is small relative to the distance between random viable metabolic networks with the same phenotype. It spans of the order of 10% of metabolic network size (circa 100 reactions). We note that this distance is an average over many and sometimes very different phenotypes, and also that it is merely an upper bound to the minimal distance between metabolic networks with different phenotypes. The reason is that we only minimized the distance between G_1 and G_2 by changing G_1. Had we changed G_2 as well we would have found even smaller minimal distances. Figure 2.4d shows how this distance depends on the number of different carbon sources a network can sustain life on. The figure shows, for phenotypes that can sustain life on increasing numbers of carbon sources (horizontal axis), the mean and standard error of the minimum distance between networks with different phenotypes. While the minimal distance increases with increasing numbers of carbon sources, this increase is small, of the order of 2% of the total genotype distance. Thus, complex constraints on metabolism do not dramatically increase the difficulty networks would encounter in evolving towards specific, novel phenotypes.

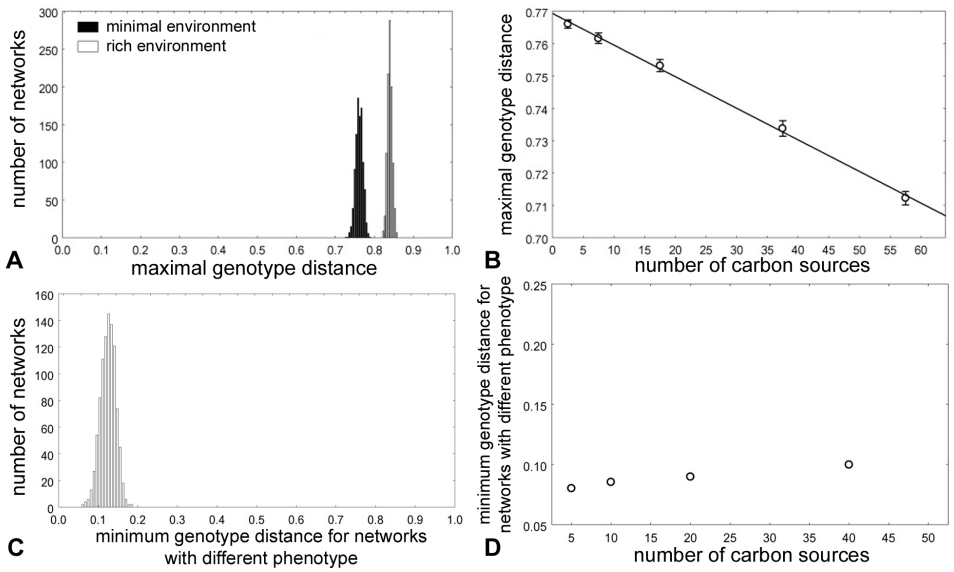

Figure 2.4: Metabolic networks with the same phenotype can have vastly different genotypes. (A) Distribution of maximum genotype distance between 1000 networks that are the end-points of random walks leading away from the initial (E. coli) network while preserving the metabolic phenotype. (B) Maximum genotype distances (vertical axis) between initial metabolic networks able to sustain life on a given number of carbon sources (horizontal axis) and 1000 final random viable metabolic networks. For each number of carbon sources 100 random walks of 10^4 mutations were carried out starting from 10 different initial networks (whiskers: 95% confidence interval). (C) The distribution of minimal genotype distance between pairs of networks with different metabolic phenotypes required to sustain life on at least one carbon source. (D) Average minimal genotype distance (the mean of the distribution in (C) as a function of the number of carbon sources. The error bars are too short to be visible in this plot.

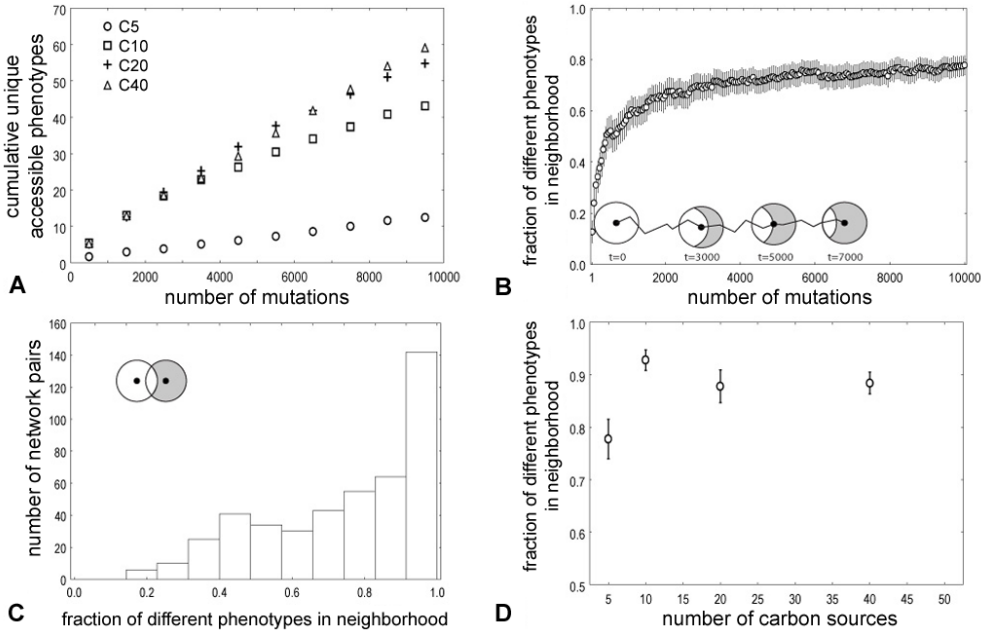

Figure 2.5: Evolving networks with conserved phenotypes can access very different novel phenotypes along their evolutionary path. (A) shows the average cumulative number of phenotypes (vertical axis) found in the neighborhood of an evolving network as a function of the number of mutations (horizontal axis) the network experienced during its evolution; (B) shows the fraction of the phenotypes in the neighborhood of the evolving network (G_t) and an initial network (G_0) that differ from one another. The diagram in the inset illustrates the increasing number of novel phenotypes in the evolving network's neighborhood (gray area of the circle) that are different from the phenotypes in the neighborhood of G_0. For pairs of random viable metabolic networks with the same phenotype; (C) shows the distribution of the fraction of different phenotypes in the neighborhoods of these networks. (D) shows the mean of the distribution (C) of phenotypic differences in the neighborhood of the network pairs versus the numbers of carbon sources they can sustain growth on. Data in (A), (B), (C) and (D) are averages over 100 random walks of 10^4 mutations starting from 10 different initial networks. In (C) only pairs of networks with the same initial network of the random walk were compared, thus 450 neighborhood comparisons. In all plots whiskers represent the 95% confidence interval.

2.4.5 Evolving networks encounter ever-new phenotypes in their immediate neighborhood

Does the genotypic plasticity of metabolic networks facilitate the discovery of novel metabolic abilities? To address this question, we examined the novel metabolic phenotypes accessible to networks that are subject to phenotype-preserving evolutionary change. By phenotypes accessible to a network, we here mean all the phenotypes that can be found in the neighborhood of this network. These are novel phenotypes that can be easily reached through a single, small genetic change. Specifically, we first carried out a random walk starting from a network with a specific metabolic phenotype, and counted the cumulative unique number of phenotypes that occurred in the neighborhood of this random walker. That is, if a phenotype occurred twice, either in the neighborhood of the same network, or in the neighborhood of a network encountered previously during the random walk, we counted it only once. Figure 2.5a shows the cumulative number of new phenotypes that such an evolving network encounters. This number does not saturate and continues to increase even though the random walk shown here comprises many thousand mutations. Second, we compared the phenotypes in the neighborhood of (i) an evolving network G_t with unchanging phenotype, and (ii) its ancestor G_0 as a function of the number of mutations t between the two networks. Specifically, we asked for the fraction of phenotypes that differ between the one-neighborhoods of the two neighborhoods. If this fraction were close to one for large t, then even two dissimilar networks might only have access to very similar metabolic phenotypes. Figure 2.5b shows, as a function of t, the fraction of different phenotypes in the neighborhood of G_0 and G_t. It is evident that this fraction approaches a large value very quickly, that is, even similar genotypes have access to a diverse spectrum of phenotypes. Third, we examined the neighborhoods of multiple end points (orange circle in Figure 2.1) of long phenotype-preserving random walks starting from the same network. Doing so tells us how different the phenotypes accessible from very different (essentially random) metabolic networks with the same phenotype are. Figure 2.5c shows the distribution of this fraction of accessible but different phenotypes for 4950 network pairs. Importantly, the vast majority of phenotypes differ among these pairs. That is, phenotypes found near one network are usually different from phenotypes near another network with the same phenotype. In sum, three independent lines of evidence show that the metabolic phenotypes accessible to networks with the same phenotype differ dramatically even for moderately different networks. Finally, we also examined how the accessibility of novel phenotypes depends on the phenotypic complexity of the evolving networks themselves, that is, on the number of carbon sources that they can support life on. In principle, all 2101 phenotypes are accessible from any metabolic genotype through a single mutation, regardless of the number of carbon sources the genotype is viable in (see

section 2.7 for a detailed explanation). However, Figure 2.5a and Figure 2.8 show that networks able to sustain life on more carbon sources encounter more novel phenotypes along their evolutionary trajectory. In addition, Figure 2.5d shows that the fraction of metabolic phenotypes that differ between the neighborhoods of random viable network pairs with the same phenotype is consistently large and shows no simple dependency on the number of carbon sources.

2.5 Discussion

Metabolic networks can evolve through the elimination of individual reactions by mutation, and through the addition of new reactions by horizontal gene transfer. We here explored a vast space of metabolic network genotypes through random changes of individual reactions that preserve a network's metabolic abilities. The ability of flux balance analysis to determine metabolic phenotypes a network's ability to sustain life in a well-defined environment containing specific carbon sources allowed us to characterize the relationship between metabolic genotypes and phenotypes. We find that metabolic networks with the same phenotype show enormous genetic plasticity, and that this plasticity aids in the evolution of novel metabolic abilities. Multiple experimental and computational studies show that a large fraction of enzyme-coding genes are dispensable in genome-scale metabolic networks. These networks continue to sustain life even upon removal of many apparently central and important reactions [49, 56, 102, 48, 106, 115, 120, 121, 122, 123]. These studies raise the question whether such robustness is an evolutionary adaptation, evolved in response to ongoing mutational pressure. Our approach of creating multiple, essentially random viable metabolic networks with pre-defined phenotypes suggests an answer to this question for the *E. coli* metabolic network. In both a glucose-minimal and a rich environment, the fraction of reactions dispensable in the *E.coli* network is not dramatically different from that of 1000 metabolic networks with the same metabolic phenotypes. This argues that the high robustness to gene deletions of *E. coli* metabolism may not be an evolutionary adaptation, but is rather typical of metabolic networks of comparable size. A caveat to this observation is that our approach allows modest fluctuations in reaction numbers (by about 14 percent) to facilitate the sampling of metabolic genotype space. These fluctuations may influence estimates of robustness by approximately the same amount. We will leave exploration of this influence to future work. Our observations go beyond preceding work which showed that a reaction's essentiality may depend on the environment [116, 124]. We demonstrate that the plasticity of metabolic networks is so great that even in a single environment, different networks with the same phenotypes may show very different essential reactions. For example, only 7.3% of all reactions essential in at least one of 1000

networks are essential in all networks. Excluding these reactions, two networks with the same phenotype differ in 74% of their essential reactions. Even in pathways as important as central energy metabolism, the vast majority of reactions are essential in only 1% of networks. One might think that networks able to thrive on many different carbon sources might show vastly more essential reactions. However, this is not the case. Reaction essentiality depends only modestly on the number of carbon sources a network can sustain life on. Gene essentiality thus strongly depends on a network's genotype, which is highly malleable. Even organisms with similar metabolic abilities may thus show very different dispensable genes in a given environment. These observations have implications for the design of antimetabolic drugs that inhibit specific metabolic reactions. Specifically, an evolutionary approach like ours may be highly useful in identifying reactions that are essential in most networks with a given metabolic phenotype, as a precursor to rationally designing drugs inhibiting these reactions. The more frequently essential a reaction is, the smaller the likelihood that a cell can circumvent it through addition or deletion [115] of other reactions. For example, the major antimetabolic antibiotics sulfonamides and trimethoprim inhibit two different reactions (dihydropteroate synthetase and dihydrofolate reductase) leading to tetrahydrofolate, an essential precursor for nucleic acid synthesis. These two reactions, however, are essential in only 40 percent of networks able to sustain life in rich medium. Figure 2.10 shows some of the ways by which nonessentiality arises in this case. Multiple bacterial species, for example, bypass the need for dihydrofolate reductase in the synthesis of nucleotide precursors, using a flavin-dependent thymidilate synthase instead [125]. A better target in the same pathway would be the enzyme dihydrofolate synthase, which our approach finds to be essential in all networks (Figure 2.10) In a similar vein, it is no coincidence that a broad class of antibiotics (penicillins, bacitracin, cephalosporins, carbapenems, vancomycin etc.) target synthesis of cell walls and membranes: Among the reactions found to be essential in all networks (Table S1), cell wall and membrane biosynthesis reactions are highly enriched. Thus, our approach lends itself to a pre-screening of metabolic reactions or reaction classes for drug targeting. Our analysis shows that vastly different networks with the same phenotype can be connected through paths of single mutations (reactions additions/deletions) in genotype space. Specifically, these paths can traverse more than three quarters of genotype space without destroying a given phenotype. This phenomenon does not depend strongly on the evolutionary constraints on a metabolic network, that is, on the number of carbon sources a network is required to sustain life on. These observations are reminiscent of genotype networks or neutral networks that have been characterized for RNA, protein, and transcriptional regulation circuits [62, 64, 126, 127, 128, 63]. In these networks, genotypes with the same phenotype form large sets in genotype

space, sets that can be connected through many single, small mutational changes. For example, proteins with the same tertiary structure and function (phenotype) often share a common ancestor, but their amino acid sequences (genotypes) have diverged beyond recognition [129, 130]. The existence of such genotype networks and the robustness it implies facilitates the evolution of new molecular functions [131, 132, 67, 65]. We here provide two lines of evidence that genotype networks may also facilitate the evolution of new metabolic phenotypes, the ability to survive on previously not utilizable carbon sources. First, we show that networks with different and arbitrary phenotypes can be found close together in genotype space. This means that from any one network, only a small fraction of genotype space needs to be traversed to find any given, novel phenotype. Second, we also analyze the neighborhood of different neutral networks with the same phenotype. This neighborhood consists of all networks that differ in only one reaction from a focal network. They are thus accessible from this network through a single mutation. We find that the neighborhoods of different networks contain very different novel phenotypes. This means that by traversing a large fraction of genotype space without changing the phenotype, one can render different novel phenotypes accessible (Figure 2.11). Put differently, even microorganisms with identical phenotypes may be able to access very different novel phenotypes. This observation points to the need to carefully choose organismal strains for engineering of novel metabolic abilities, such as the production of biofuels, or the degradation of toxic compounds in bioremediation. The right choice may mean that only a small alteration, such as the addition of one reaction to a metabolic network, is sufficient to produce a desired new phenotype. Consider the example of the carbon source melibiose, a sugar similar to lactose and made of the same two monosaccharides (galactose and glucose) but differing in the glycosidic link between them. While lactose can be metabolized by many microbes, melibiose is a less commonly utilizable compound. The metabolization also requires different enzymes (α-galactosidase for melibiose and β-galactosidase for lactose). The metabolic ability to use melibiose is desirable, for example in yeast, where cells have been engineered to utilize melibiose to improve efficiency and reduce waste in fermented dairy products [133]. Among the networks with identical metabolic phenotypes that we examine, there are networks where adding the α-galactosidase reaction is sufficient to endow the network with melibiose utilization. In contrast, in other networks with the same phenotype the addition of this reaction is not sufficient (even though both networks are able to grow on glucose). The reason is that these latter networks are unable to excrete the excess galactose from the degradation of melibiose. Another example involves the addition to a network of a single reaction catalyzing the transfer of a phosphor group from a phospho-histidine to galactitol. This reaction produces galactitol 1-phosphate, and it enables the network to grow on galactitol. In another network

with the same phenotype, the addition of this reaction does not have the same result. The reason is that the first network contains other reactions that enable it to convert of galactitol 1-phosphate into galactose, which it can grow on. We next motivate the choice of metabolic network sizes for our work. Flux balance analysis has been used to show that a significant number of reactions in *E. coli*, when removed, show no impact on optimal growth in several different environments [134]. This observation might lead one to suppose that phenotype-preserving paths through genotype space are long merely because many reactions are never essential. However, this is not the case. For example, although the fraction of essential reactions in *E. coli* is merely 28% when considering a glucose minimal environment, this fraction rises to 43% when considering growth on each of the more than 81 carbon sources we examined here. In addition, when considering the influence of the genetic background, we observe that 66% of the reactions appear as essential in at least one of the many randomized viable metabolic network in a glucose minimal environment, and 81% of reactions become essential when we consider the full spectrum of 81 carbon sources. This fraction of essential reactions would undoubtedly have risen further if we had the computational means to analyze additional carbon sources and genetic backgrounds. Taken together, these observations mean that essentiality of reactions depends on environment and genetic background, and that there may not be a meaningful reduced reaction set that is always under selection. These observations, and our desire to compare properties of our sampled networks to the *E. coli* network prompted our choice of network size. Flux balance analysis has limitations in how precisely it can predict growth or by-product secretion after gene knockouts [97], which may depend on the choice of optimization principle [135] and flux maximization method [49]. These limitations are connected to how metabolic genes are regulated, and they do not affect our study because we are not concerned with regulatory evolution. For our purposes, it is sufficient to evaluate if an organism represented by a metabolic network is viable in principle, based on the complement of enzymes it carries and the biomass precursors it can synthesize given a spectrum of nutrients. The potential problem of limited and likely biased information about the set of biochemical reactions that occur in nature does not affect our results qualitatively. The reason is that any increase in the number of known biochemical reactions will cause the appearance of alternative pathways, lowering the number of essential reactions, and thus increasing the robustness and the plasticity of metabolic networks. Aside from these caveats, the biggest limitation of the approach presented here lies in its computational demands. Determining the metabolic phenotypes of networks in the neighborhood of a single genome-scale network for 101 carbon sources requires the solution of 5.85×10^5 ($= 101 \times 5800$) complex linear programming problems [97]. For our simulations we analyzed more than 20 000 such genomes and this was cur-

rently at the limit of computational feasibility. This limitation will undoubtedly be ameliorated with time. In sum, the approach proposed here can provide various insights into the organization of metabolic networks. It demonstrates that the architecture of such networks shows high plasticity, even for single environments, a property that facilitates the evolution of new metabolic functions. It suggests a method to target metabolic reactions for rational drug design, and shows that the plasticity of metabolic networks creates both opportunities and constraints for the evolution of novel metabolic abilities.

2.6 Methods

2.6.1 Random walks in genotype space

We explore the vast space of metabolic networks by long random walks that leave a network's ability to synthesize all essential biomass components unchanged. Each step of the random walks we use has two parts. The first part consists of mutation, the deletion of a randomly chosen reaction from a network, or the addition of a new randomly chosen reaction from the global reaction set above. We constrain variation in the number of reactions in this random walk by means of a bias in the choice of mutation that depends linearly on the number of reactions in the metabolic network (see section 2.7). With this procedure, the networks have always approximately 1000 reactions throughout the simulations. In the second part of a random walk's step, we apply flux balance analysis to verify that the new metabolic network still has the same phenotype, i.e., that it can still grow on the same specific set of carbon sources. If so, the mutated network is accepted and the next step of the walk starts with the mutated network; if not, the mutated network is rejected, and the next step of the random walk starts with the previous (unmutated) network. Methods are described in greater detail in the section 2.7.

2.7 Supplementary Text

2.7.1 Flux balance analysis

Flux balance analysis (FBA) is a computational approach to characterize the behavior of large ($> 10^3$ reactions) chemical reaction networks [47, 97, 51, 136, 112]. In FBA, a network is represented by a set of stoichiometric equations describing chemical reactions. FBA takes advantage of the invariance of metabolite concentrations in a metabolic network that is in steady-state. This invariance implies that only some distributions of metabolic fluxes rates at which individual reactions

proceed do not violate the law of mass conservation. Among these allowed steady-state fluxes, FBA can identify fluxes that have particular properties of interest in a given environment, defined by a maximum influx of external nutrients. We are here interested in one key property, namely whether a given metabolic network can sustain life in a given environment. That is, can it produce all key biochemical precursors necessary to sustain growth and energy production? Flux balance analysis allows us to answer this question. In our work we use a set of biochemical precursors from *E. coli* [137, 138, 53] as the set of required compounds a network needs to synthesize, by using linear programming to optimize the flux through a specific objective function, in this case the reaction representing the production of biomass precursors we are able to know if a specific metabolic network is able to synthesize the precursors or not. The precursors include all 20 proteinaceous amino acids, nucleotides, deoxynucleotides, putrescine, spermidine, 5-methyltetrahydrofolate, coenzyme-A, acetyl-CoA, succinyl-CoA, cardiolipin, FAD, NAD, NADH, NADP, NADPH, glycogen, lipopolysaccharide, phosphatidylethanolamine, peptidoglycan, phosphatidylglycerol, phosphatidylserine and UDPglucose. Flux balance analysis relies on linear programming [55] to identify network properties of interest. We here used the packages CPLEX (11.0, ILOG; http://www.ilog.com/) and CLP (1.4, Coin-OR; https://projects.coin-or.org/Clp) to solve the associated linear programming problems. We studied metabolic networks in one main aerobic environment, a minimal environment composed of one or more carbon sources, oxygen, ammonia, inorganic phosphate, sulfate, sodium, potassium, iron, protons and water. When studying different growth phenotypes of a particular metabolic network we here focus on carbon sources, and thus vary only the carbon source in this minimal aerobic environment. For example, when we say that a network is able to sustain life on five specific carbon sources, we mean that it produces all essential biosynthetic precursors (a non-zero growth flux) when each of these carbon sources is provided as the sole carbon source in a minimal medium. This implies of course that any subset or combination of these five carbon sources would also suffice to sustain life. The 101 possible carbon sources we study here represent a tiny fraction of 101/550=18.3% of all carbon-containing metabolites in E. coli, and an even smaller fraction 101/4425=2.2% of carbon containing metabolites in our universe of metabolites (Figure 2.6a). Many metabolites other than those from E. coli can and do serve as carbon sources for other prokaryotes. Computational limitations prevented us from analyzing more complex carbon phenotypes. For some analyses, we also used a rich aerobic environment [113]. This environment is composed of 36 metabolites, which includes the proteinaceous aminoacids , carbon dioxide, thiamin, nicotinamide mononucleotide, pantoate, and all the metabolites available in the minimal environment.

2.7.2 The global reaction set

Each metabolic network is a point in a much larger genotype space of networks. For the universe of reactions that can occur in these networks we used data from the LIGAND database [119] of the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.ad.jp/kegg/ligand.html) [50]. The LIGAND database is a database of chemical compounds and reactions in biological pathways that is compiled from pathway maps of metabolism of carbohydrates, energy, lipids, nucleotides, amino acids and others. Also included in the database is the list of recommended names for enzymes given by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (http://www.chem.qmul.ac.uk/iubmb/enzyme/) which includes all categorized enzymes (oxidoreductases, transferases, hydrolases, lyases, isomerases and ligases). We used specifically the REACTION and COMPOUND sections of the LIGAND database to construct our global reaction set. From this data we pruned (i) all reactions involving general polymer metabolites of unspecified numbers of monomer units ($C_2H_6(CH_2)_n$), or, similarly, general polymerization reactions that were of the form $A_n + B \rightarrow A_n + 1$, because their abstract form makes them unsuitable for stoichiometric analysis, (ii) reactions involving glycans, because of their complex structure, (iii) reactions that were not stoichiometrically or elementally balanced, and (v) reactions involving complex metabolites without chemical information about their structure. The starting point of our work is the *E. coli* metabolic network (iJR904) [95] which comprises 726 reactions (excluding transport reactions). We merged all reactions in the *E.coli* network with the reactions in the KEGG dataset. (Only few E.coli reactions, such as specific nutrient or waste transport reactions necessary for FBA, and some specific polymerization reactions were not already in the KEGG database.) After these steps of pruning and merging, our global reaction set consisted of 5870 reactions and 4634 metabolites.

2.7.3 The set of networks able to sustain life on a given set of carbon sources is connected

We note that two network genotypes able to sustain life on a given set of carbon sources can be reached from one another through single mutations in genotype space without abolishing this ability. To see this, consider the set of reactions R_1 and R_2 that occur in two arbitrary such networks. Denote the network formed of the union of these reaction sets as $R_1 \cup R_2$. Note that the addition of a chemical reaction to any network will not abolish its ability to sustain life on any given spectrum of carbon sources. This means that there exists a sequence of single reaction changes $(\mu_1, \ldots, \mu_n)$ that leads from R_1 to $R_1 \cup R_2$, as well as another sequence $(\nu_1, \ldots, \nu_m)$ that leads from R_2 to $R_1 \cup R_2$. Denote for any mutational

change ν its opposite as $\overline{\nu}$. That is, if ν is the deletion of a reaction r, then $\overline{\nu}$ is the addition of the same reaction to a network that does not contain it, and vice versa. It follows from the above considerations that the sequence of mutations $(\mu_1, \ldots, \mu_n, \overline{\nu_1}, \ldots, \overline{\nu_m})$ lead from R_1 to R_2 without abolishing the ability to sustain life.

2.7.4 Random walks in genotype space

We explore the vast space of metabolic networks by long random walks that leave a network's ability to synthesize all essential biomass components unchanged. In each step of such a walk, one reaction is eliminated or added to a network. During a sufficiently long random walk, the reactions in a network become effectively randomized, yet the phenotype remains constant. We are well aware that recombination through unequal cross-over or horizontal gene transfer may change more than one reaction at a time, but we focus here on individual reactions, because they are the smallest sensible unit of change. In biological evolution, natural selection probably plays a major role in changing the structure of biological networks. For example, the addition of a reaction to a metabolic network may become favorable in a new environment, and go to fixation without affecting the network's ability to sustain life in the original environment. Because the detailed modeling of these and similar evolutionary dynamics would require us to make many ad hoc assumptions, we instead focus on the more tractable question whether changes can preserve metabolic phenotypes. Each step of the random walks we use has two parts. The first part consists of mutation, the deletion of a randomly chosen reaction from a network, or the addition of a new randomly chosen reaction from the global reaction set above. We constrain variation in the number of reactions in this random walk by means of a bias in the choice of mutation that depends linearly on the number of reactions in the metabolic network. Specifically, the probability that a reaction is deleted (as opposed to added) is given by $p_{del} = R/R_0 - 0.5$, where R is the number of reactions in the current network, and R_0 is the number of reactions in the initial network, i.e., at the start of the random walk. With this procedure the networks have approximately 1000 reactions throughout our random walks, because we used the *E. coli* network as the starting network for these random walks. Without constraint, the number of reactions in a metabolic network would steadily increase, because networks with more reactions are more likely to sustain life in a given environment. We note that our approach allows an increase in the number of reactions of roughly 14 percent relative to the starting (E.coli) network. It would thus not bias our estimates of the robustness of randomized viable networks by more than that amount. In the second part of a random walk's step, we apply flux balance analysis to verify that the new metabolic network still has the same phenotype, i.e., that it can still grow on a specific set of carbon sources. If

so, the mutated network is accepted and the next step of the walk starts with the mutated network; if not, the mutated network is rejected, and the next step of the random walk starts with the previous (unmutated) network. In carrying out these random walks, it is important to proceed for as many steps as are needed to "erase" the "memory" of the initial state. To arrive at a heuristic criterion for the required number of steps, we determined, first, the autocorrelation function of the growth flux [97, 95] along a random walk. This autocorrelation function decays to a value of zero in around 500 (Figure 2.6a) mutational steps. Unless otherwise mentioned, the number of mutational steps we use in our analysis is 10^4, and thus vastly exceeds this required number of steps. Second, we recorded the (Hamming) distances of the random walker to the initial network during random walks. This distance first increases, and then reaches a stochastic equilibrium after about 5000 steps, a number smaller than the 10^4 steps we routinely used. Finally, we note that the networks we studied have less than 10^3 reactions. In a random walk of 10^4 steps, each reaction is thus mutated many times over. Taken together, these observations show that 10^4 steps are more than sufficient to effectively randomize the initial network. We will refer to the end-point of such a random walk as a random viable metabolic network with a given phenotype. It may be very different from a random sample of chemical reactions from the whole set of reactions we consider (Figure 2.6a), which may not sustain life in any environment. We call the random walk defined above an unbiased random walk, because it does not lead into a particular direction. To study different aspects of network evolution, we also use several random walks with the following specific biases. First, to study the diameter of the set of genotypes with a given phenotype, it is necessary to obtain metabolic networks whose Hamming distance to the starting network is as large as possible. To this end, we used a forced random walk. Here, whenever a reaction that occurred in the initial network is removed from the network, we do not allow it to be added again. In this manner, no individual step of the walk can decrease the Hamming distance to an initial network. Second, to obtain networks that grow on a specific target number k_T of carbon sources (without regard to the identity of these carbon sources), we start with a network that sustains growth on some number k0 of carbon sources. If $k_0 > k_T$, we allow only mutations that maintain or decrease the number of carbon sources a network is able to grow on. Specifically, we revert a newly mutated network to its previous state whenever the number of carbon sources that it grows on is greater than the previous state, or smaller than the target number of carbon sources. If $k_0 < k_T$, we allow only mutations that maintain or increase the number of carbon sources a network is able to grow on. Third, to find a network that grows on a specific target set of carbon sources $(C_0, C_1, \ldots, C_k)$, i.e., a network whose phenotype P_T is a specific binary vector (Figure 2.6), we accept new mutations only when they decrease or do

not alter the Hamming distance between the current phenotype P and the target phenotype P_T , $d(P, P_T)$.

2.7.5 Characterizing maximum genotype distances

To study the maximal distances of two genotypes with the same phenotype, we began with the *E. coli* network, and first obtained one network expressing different phenotypes distinguished by the number $k = 5, 10, 20, 40$ of carbon sources they grow on. From each of these initial networks, we performed 100 forced random walks of 10^4 mutational steps each that conserved the phenotype of the initial network. We then recorded the distribution of the Hamming distances between the genotype G_0 of each starter network and the maximally distant network G_T at the end of the random walk, and studied the properties of this distribution as a function of k.

2.7.6 Characterizing minimum genotypic distances for networks with different phenotypes

To characterize the minimal genotype distance that separates a pair of genotypes (G_1, G_2) with different phenotypes (P_1,P_2), we performed the following analysis. For each class of phenotypes that grow on $k = 5, 10, 20, 40$ carbon sources, we generated 100 pairs of random viable metabolic networks. Each network in a pair has a different (random) phenotype, with the constraint that both networks from a pair can sustain life on the same number of carbon sources. For each pair, we then performed a forced random walk of 1000 steps, beginning with the first network G_1 (leaving the genotype G_2 of the second network unchanged). Each mutation in this random walk was required to (i) keep the network's phenotype unchanged, and (ii) not increase the Hamming distance to G_1. We recorded the minimal distance encountered in this random walk.

2.7.7 Phenotype accessibility is independent of the number of carbon sources the metabolic network is viable in

In our simulations we always considered phenotypes based on the full spectrum of 101 carbon sources. When we perform a random walk for a network that grows on 20 carbon sources, we only allow mutations to be accepted if they leave unchanged the phenotype (neither introduce the ability to be viable in an additional carbon source, or lose the viability in one of the original 20 carbon sources). However when we check the diversity of the phenotypic neighborhoods we do not limit new phenotypes to 20 carbon sources. Instead, we consider all the 2^{101} phenotypes that

101 carbon sources allow. A network that is viable in one carbon source may have a mutant that through a reaction addition will have a phenotype viable in, for example, 20 carbon sources. In the same manner, a network viable in 20 carbon sources may, through a reaction deletion, become viable in only one carbon source. This merely serves to show that all 2^{101} phenotypes are accessible in principle in all our analyses. In other words, a larger number of carbon sources does not enable more phenotypes

2.8 Supplementary Figures

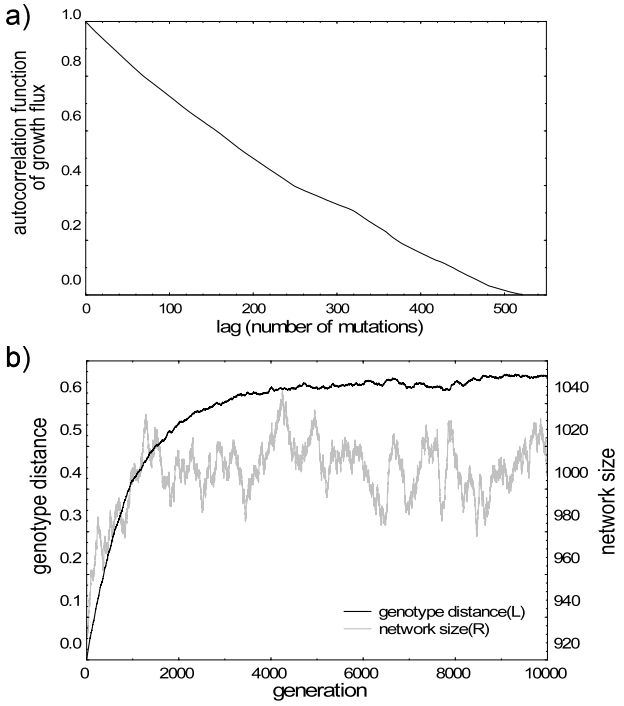

Figure 2.6: Random walks in genotype space. a) Autocorrelation function of growth flux in an unbiased random walk of 10 000 generations starting from the E. coli metabolic network. The autocorrelation function was calculated for the last 5 000 generations. b) A sample trajectory of a random walk starting from the E.coli metabolic network, showing both the number of reactions in the evolving network, as well as the genotype distance (normalized Hamming distance) between the evolving network and the initial network. When the genotypes of both networks are represented by binary vectors indicating the presence or absence of reactions (see Figure 2.1a), the normalized Hamming distance corresponds to the fraction of entries in these two vectors that are different.

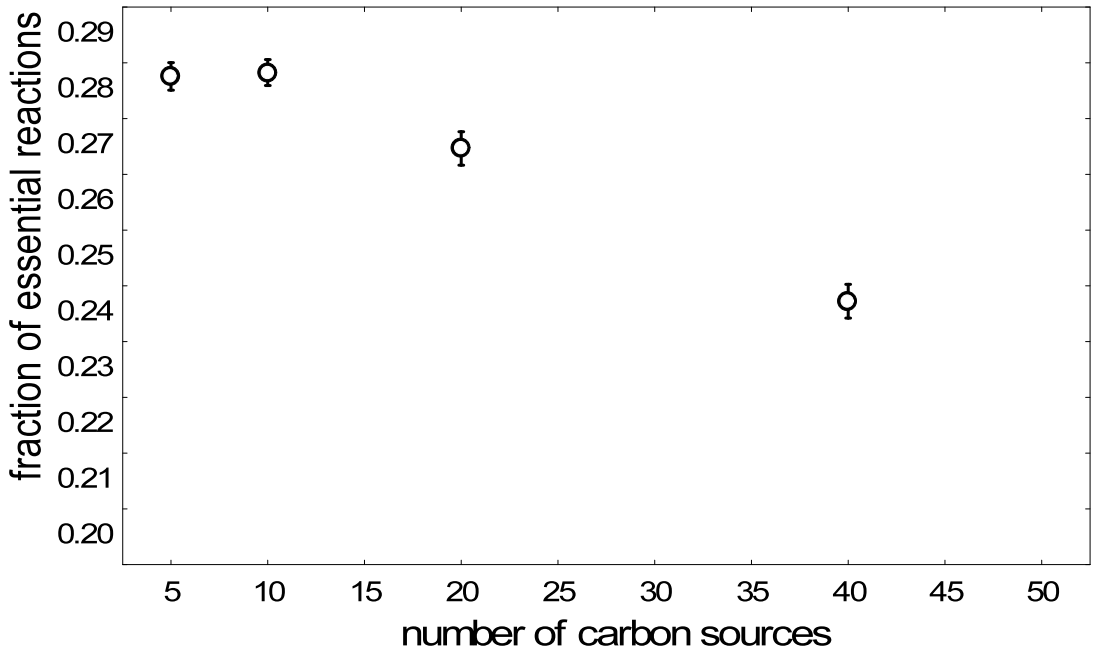

Figure 2.7: The fraction of reactions essential in a complex environment decreases with environmental complexity. Average fraction of essential reactions (vertical axis) as a function of the number of carbon sources a network can sustain life in (horizontal axis). A reaction is called essential here, if it is essential in an environment that contains all of the carbon sources a network is required to grow on. For each number of carbon sources 10 different initial networks were generated, as described in Methods, and for each of these 10 networks 10 random walks were carried out. Each circle on the plot is thus based on 100 networks (whiskers: 95% confidence interval). See Methods for details on how the initial networks were generated.

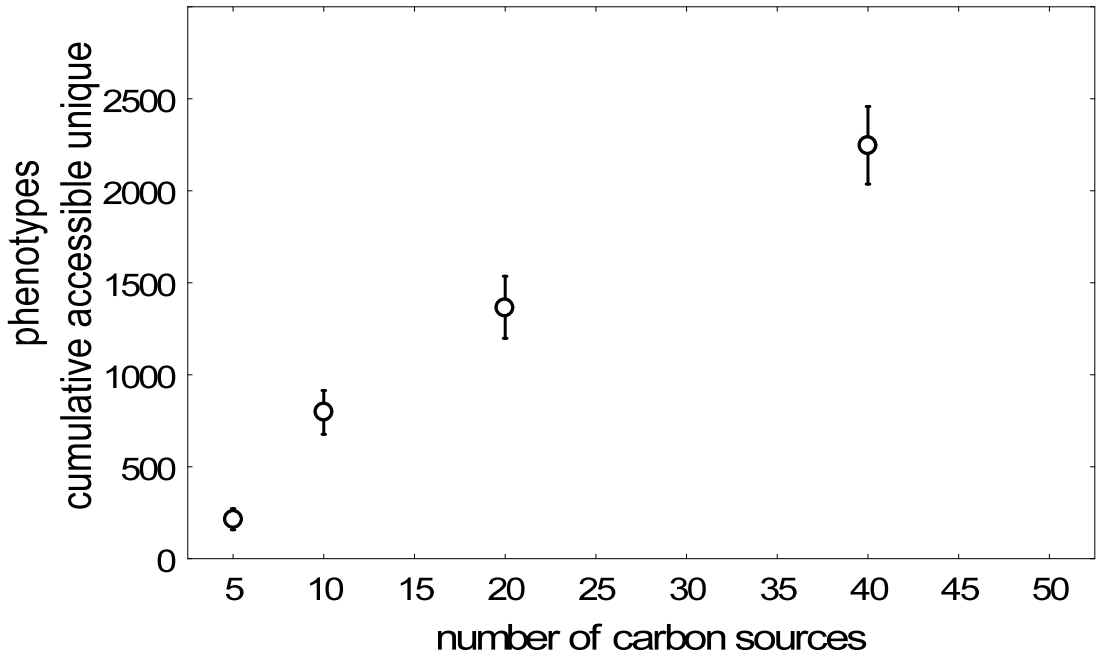

Figure 2.8: Networks that can grow on more carbon sources encounter more novel phenotype during their evolution. The average cumulative number of phenotypes (vertical axis) found in the neighborhood of an evolving metabolic network at the endpoints of 100 phenotype-preserving random walks is shown as a function of the number of carbon sources the initial networks can grow on. For each number of carbon sources shown, the data is an average over 10 independently generated initial networks, and over 10 random walks starting from each of these 10 networks.

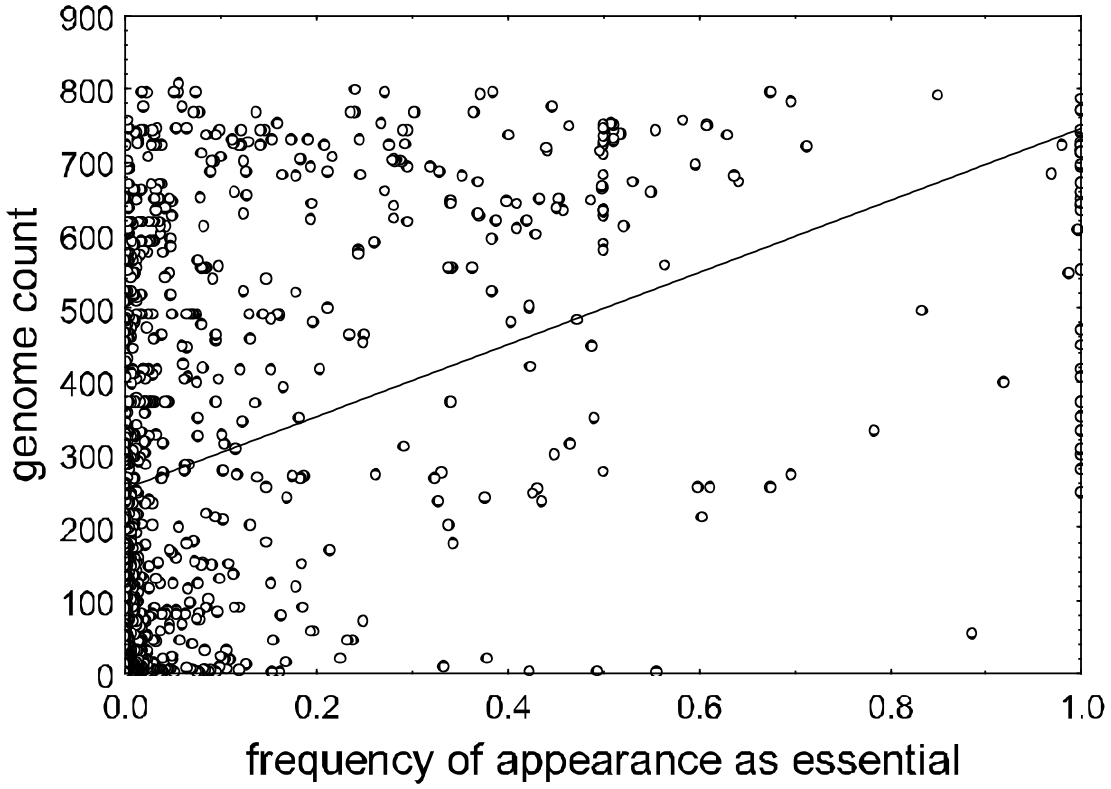

Figure 2.9: Reaction essentiality and gene appearance in prokaryotic genomes. Correlation of frequency of reaction essentiality in random metabolic networks and number of genomes carrying an enzyme-coding gene catalyzing that reaction. Pearson's r = 0.45; $p = 2.2 \times 10^{16}$. This analysis uses enzyme-coding genes from 875 prokaryotic genomes in the KEGG database.

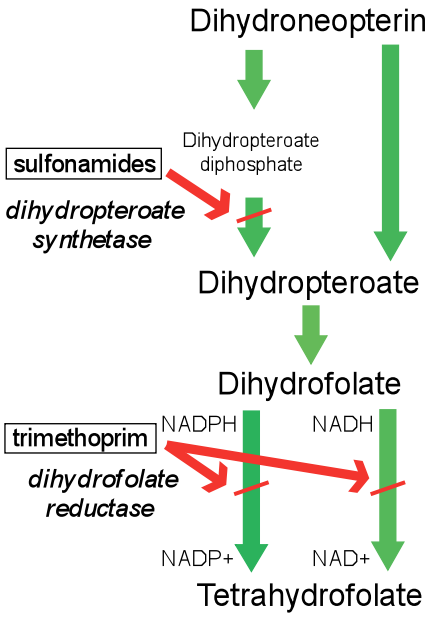

Figure 2.10: Reactions in tetrahydrofolate biosynthesis and their essentiality. We found that the reaction dihydropteroate synthetase, a target of sulfonamides, is essential in 41% of the metabolic networks we studied, while the other reaction producing dihydropteroate is essential in 56.1% of networks. In the remaining 2.9% of networks, both reactions appear, but none are essential. These observations have a straightforward explanation. Dihydropteroate is an essential metabolite. Because only two alternative reactions exist to make dihydropteroate, whenever one of these reactions is missing, the other is an essential reaction. Whenever both reactions are present, neither reaction is essential. For the production of tetrahydrofolate from dihydrofolate, there exist, similarly, two parallel dihydrofolate reductase reactions. These reactions are the target of trimethoprim. The reactions are only distinguished by the molecule that acts as the electron donor, either NADH or NADPH. Individually, these reactions appear as essential in only 30%40% of networks. In addition, only 66.2% of networks cannot tolerate the removal of both reactions. The reason is that there are alternative paths (not shown) that bypass the direct production of tetrahydrofolate from dihydrofolate.

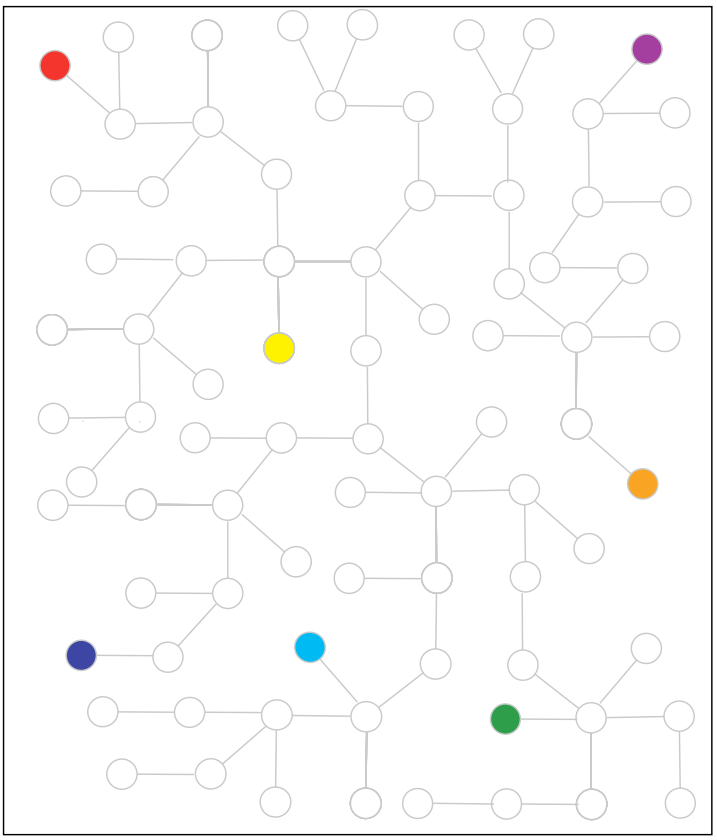

Figure 2.11: The connectedness of metabolic networks with the same phenotype facilitates access to new metabolic phenotypes. The rectangle symbolizes genotype space, and the grey circles symbolize metabolic networks with a given metabolic phenotype. The colored circles stand for metabolic networks with a novel phenotype. Different novel phenotypes (different colors) are accessible from different networks (points) in genotype space with the same phenotype.

3 Genotype networks, innovation and robustness in sulfur metabolism

João F. Matias Rodrigues and Andreas Wagner
[*submitted*, **2010**]

3.1 Abstract

A metabolism is a complex network of chemical reactions. This network synthesizes multiple small precursor molecules of biomass from chemicals that occur in the environment. The metabolic network of any one organism is encoded by a metabolic genotype, defined as the set of enzyme-coding genes whose products catalyze the network's reactions. Each metabolic genotype has a metabolic phenotype. We define this metabolic phenotype as the spectrum of different sources of a chemical element that a metabolism can use to synthesize biomass. We here focus on the element sulfur. We study properties of the space of all possible metabolic genotypes in sulfur metabolism by analyzing random metabolic genotypes that are viable on different numbers of sulfur sources. We show that metabolic genotypes with the same phenotype form large connected genotype networks — networks of metabolic networks — that extend far through metabolic genotype space. How far they reach through this space depends linearly on the number of super-essential reactions. A super-essential reaction is an essential reaction that occurs in all networks viable in a given environment. Metabolic networks can differ in how robust their phenotype is to the removal of individual reactions. We find that this robustness depends on metabolic network size, and on other variables, such as the size of minimal metabolic networks whose reactions are all essential in a specific environment. We show that different neighborhoods of any genotype network harbor very different novel phenotypes, metabolic innovations that can sustain life on novel sulfur sources. We also analyze the ability of evolving populations of metabolic networks to explore novel metabolic phenotypes. This ability is facilitated by the existence of genotype networks, because different neighborhoods of these networks contain very different novel phenotypes. We show that the space of metabolic genotypes involved in sulfur metabolism is organized similarly to that of carbon metabolism. We demonstrate that the maximum genotype distance and robustness of metabolic networks can be explained by the number of superessential reactions and by the sizes of minimal metabolic networks viable in an environment. In contrast to the genotype space of macromolecules, where phenotypic robustness may facilitate phenotypic innovation, we show that here the ability to access novel phenotypes does not monotonically increase with robustness.

3.2 Introduction

In any biological system, genotypes contain the information needed to make phenotypes. The relationship between genotype and phenotype is also known as a genotype-phenotype map [60]. The ability to analyze different kinds of biological systems computationally has allowed a detailed characterization of genotype-

phenotype maps for different systems. One common feature of genotype-phenotype maps is the existence of genotype networks, connected sets of genotypes that adopt the same phenotype. They exist in systems as different as model proteins [61], RNA secondary structures [62], regulatory circuits [63], and metabolic networks [139, 140]. Another feature is the large phenotypic diversity that is found in different neighborhoods of a genotype network [62, 63, 139, 140]. These two properties facilitate the exploration of novel and potentially beneficial phenotypes in genotype space. By analyzing genotype-phenotype maps of different systems, one can identify general features of genotype maps, as well as features that are specific to a system.

In this work we concentrate on the genotype-phenotype maps of metabolic networks involved in the utilization of sulfur. We have two aims. Aim 1 is to examine how general earlier observations about the genotype-phenotype map of carbon metabolism are [139, 140]. We do so by examining if these observations also apply to sulfur metabolism. In particular, we investigate the existence of genotype networks whose members share the same phenotype, and the amount of phenotypic diversity in their neighborhoods. Aim 2 is to study how rapidly evolving *populations* of networks "discover" metabolic innovations in metabolic genotype space. Specifically, we are interested in how the rate of discovery depends on the robustness of a metabolic system. This robustness indicates a metabolic network's ability to preserve its biosynthetic capacity upon random removal of reactions. Previous work on macromolecules showed that the robustness of a molecule's phenotype to mutations can accelerate the rate at which evolving populations discover new phenotypes [141, 74]. We will ask whether the same holds for metabolic systems.

Carbon metabolism comprises so many reactions that the computational demands of studying population processes in its genotype space are too high for current computational technology. Sulfur metabolism, in contrast, comprises a smaller number of chemical reactions, which renders the computational analysis of population processes more tractable. Despite being involved in fewer reactions, sulfur is no less essential to biological organisms than other elements, such as carbon or nitrogen. Sulfur is a versatile and integral element in the biochemistry of organisms [142, 143]. Its presence in biological organisms ranges from 0.5% to 50% of dry weight [142]. It occurs in multiple oxidation states, ranging from the highly oxidized S^{4+} to the reduced state S^{2-}. This versatility in oxidation state may explain the diversity of sulfur metabolism and why it is involved in both anabolism as well as catabolism. In catabolism, depending on the environment, sulfur can be used as an electron acceptor or an electron donor, and in some cases even both as donor and acceptor. In anabolism, sulfur must first be reduced in a sequence of energetically expensive steps before being incorporated into biomass [142].

Sulfur is present in two major constituents of biomass, the amino-acid cysteine,

which confers stability to proteins through disulfide bonds, and the amino acid methionine, which is the first amino acid of many proteins. Sulfur is also a part of S-adenosylmethionine (also known as AdoMet or SAM). This compound is a cysteine metabolite that is a major methyl donor to the methyl carrier metabolite tetrahydrofolate, which is indispensable for amino acid synthesis, and for the methylation of biomolecules. Furthermore, sulfur is the active element in coenzyme-A, an acyl carrier metabolite involved in the calvin cycle and in lipid synthesis. Sulfur is also present in the active core of iron-sulfur proteins, which are involved in a number of important reactions. Examples include nitrogenase, which enables the fixation of nitrogen, and hemoglobin, which enables the transport of oxygen. Another prominent molecule involving sulfur is glutathione, a peptide responsible for protection against oxidative stress in cells.
We next outline the order of our analyses in the Results section. First we introduce two concepts that allow us to estimate some properties of genotype space organization. The first is that of a minimal metabolic network. This is a metabolic network from which no reactions can be removed without destroying its viability in a given environment, that is, its ability to synthesize all essential biomass molecules. The second concept is that of a superessential reaction. For our purpose, a superessential reaction is an essential chemical reaction that occurs in all minimal networks. After these preliminary analyses, we demonstrate the existence of long phenotype-preserving paths through metabolic genotypes space that allow exploring this space through many single phenotype-preserving mutations (aim 1). The maximum length of these paths and metabolic network size can be estimated and varies linearly with the number of superessential reactions. We show that the robustness of metabolic networks depends both on their size and on the average size of minimal metabolic networks viable on a given number of sulfur sources. Next, we show that the existence of neutral paths allows evolving metabolic networks to encounter an increasing number of novel phenotypes (aim 2). We finally explore the relationship between robustness and a population's ability to access novel phenotypes through changes in a network's reactions. In contrast to macromolecules, where robustness may facilitate phenotypic innovation [141, 74], we find that the ability to find novel phenotypes in our system peaks at intermediate robustness.

3.3 Results

3.3.1 The model

We follow an approach taken in a previous study of large-scale metabolic networks [139]. We define a metabolic *genotype* as the set of biochemical reactions that

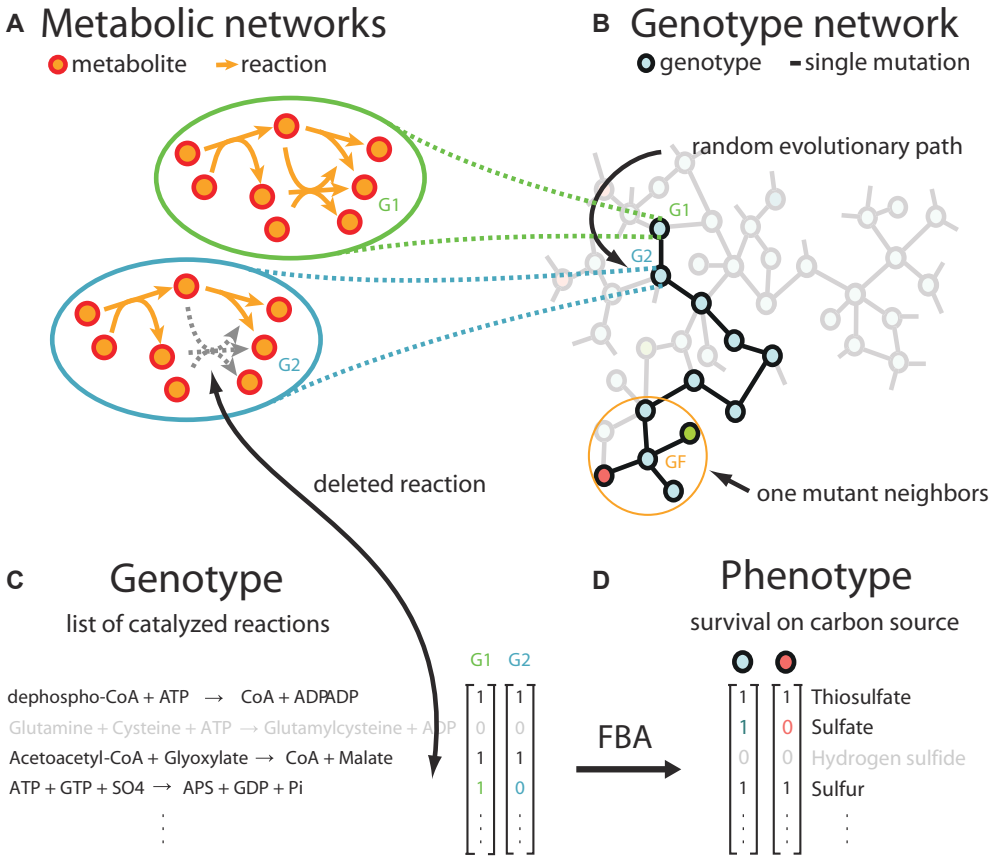

Figure 3.1: Genotype-phenotype map of metabolic networks. Different representations of a hypothetical metabolic network (A), as a node in a genotype network (B), or as a binary vector (C) listing the reactions in the network. Each genotype (circles) on the genotype network in (B) has 1221 neighbors (not all edges are drawn) that differ by a single mutation. Neighbors in (B) are connected by edges. The colors of the genotypes represent different phenotypes. The phenotypes of the metabolic networks are computed using flux balance analysis applied to 124 environments with different sulfur sources. Two hypothetical phenotypes are represented in (D) as binary vectors listing the environments a genotype is viable in (D). Random evolutionary walks can be seen as paths on a genotype network that stay on genotypes with the same phenotype (represented as the genotype color). "Mutations" correspond to additions or deletions of individual reactions from the metabolic network. The number of genotypes in the genotype space is 2^{1221}.

may take place in an organism, and that are catalyzed by gene-encoded enzymes. The set of all reactions used in this work is a subset of 1221 reactions out of 5871 reactions we curated previously [139] from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [50]. These reactions comprise all elementally-balanced reactions that involve sulfur containing metabolites (see methods section 3.6 for details). A metabolic genotype can be represented in at least 2 different ways (Figure 3.1). The first views it as a metabolic network graph whose nodes are metabolites. Reactions are represented as directed links from substrate metabolites to product metabolites (Figure 3.1A). The second views it as a list of reactions (Figure 3.1C), or, equivalently a binary vector whose length — 1221 reactions in our case — corresponds to the number of reactions in a known reaction "universe". Each position i in this vector corresponds to a reaction. Its values ('0') or ('1') at position i indicate the inability or ability of the organism to catalyze the corresponding reaction (Figure 3.1C). We define the *phenotype* of a metabolic network as the subset of sulfur sources (out of 124 possible sources we consider, see methods section 3.6) that allow the network (metabolic genotype) to synthesize all biomass components, if one of the sulfur sources is provided as the sole sulfur source to the organism. We represent this phenotype as a binary vector of length 124 whose entry at position i indicates viability if sulfur source i is the sole sulfur source (Figure 3.1D). This is not the only way to define a metabolic phenotype, but it is appropriate for our purpose. An obvious alternative phenotype definition would count the number of biomass metabolites that a network can produce in a given environment. However, because all these metabolites are essential for survival of an organism, networks that cannot synthesize some of them are of limited biological relevance. Additional advantages of the phenotype definition we chose are that it allows a straightforward and systematic comparison of phenotypes, and enables us to study metabolic innovation in a biologically sensible way. Using this phenotype definition, a metabolic innovation is the ability to synthesize biomass metabolites from a new, previously unusable sulfur source. To determine metabolic phenotypes from genotypes, we use flux balance analysis [46], a computational method that finds a growth-maximizing steady-state metabolic flux through all reactions in a metabolic network. This method requires information about the stoichiometry of every metabolic reaction, a maximally allowed flux of each metabolite in and out of the environment, and information about an organism's biomass composition (see methods section 3.6 for details). We focus on a metabolic network's qualitative ability to produce all sulfur-containing biomass precursors. We will study networks that are able to do so from each one of a specific set of sole sulfur sources. For brevity, we call such networks *viable*. We will also refer to the number S of sulfur sources that a metabolic network must be viable on as the *environmental demand* imposed on the network. We next introduce the concept of a genotype

network for metabolic networks (Figure 3.1B) [139]. The nodes in this network correspond to individual genotypes (metabolic networks) *with the same phenotype*. Two genotypes are linked — they are *neighbors* — if they differ in a single reaction. A genotype network thus is a network of metabolic networks. This concept is useful when we examine the evolution of metabolic networks through the addition and elimination of metabolic reactions, which can occur, for example, by horizontal gene transfer [17, 99], or through loss-of-function mutations in enzyme-coding genes. Consider the metabolic network genotype G_1 of some organism. This genotype is a node on the genotype network associated with this genotype's phenotype. If some variant G_2 of this network — obtained through an addition or a deletion of a reaction — has the same phenotype as G_1, it will be a neighbor of G_1 on the same genotype network. In this manner, one can envision phenotype-preserving evolutionary change of metabolic genotypes as a path through a genotype network. Such paths correspond to successive hops from genotype to genotype, by way of the edges connecting neighboring genotypes (Figure 3.1B). For our analysis, it will be useful to define a *distance* D between two metabolic network genotypes as the fraction of reactions in which two metabolic networks differ, or

$$D = 1 - \frac{2R_C}{N_1 + N_2}, \tag{3.1}$$

where R_c is the number of reactions shared by both networks, and N_1 and N_2 are the total numbers of reactions in networks G_1 and G_2, respectively. This formula simplifies to $D = 1 - R_C/N$ when both networks have the same size N.

3.3.2 Minimal viable metabolic networks can be diverse and contain many superessential reactions.

We begin with an analysis of minimal viable networks, which provides insights into the reactions that are essential to utilize a specific set of sulfur sources. A minimal metabolic network is a network in which all reactions are essential and none can be removed without rendering it inviable. For any one given phenotype P, there may be multiple viable minimal networks. Random minimal networks can be generated by starting from a network comprised of all 1221 reactions — which is viable on all sulfur sources — and eliminating randomly chosen reactions one-by-one, until no reactions can be further removed without rendering the network inviable on the sulfur sources defined by P. We note that a minimal network is not the same as the network with the smallest possible number of reactions with a given phenotype, which could be very difficult to find in a vast metabolic genotype space. We generated 1000 random minimal metabolic networks viable on a given number S of sulfur sources (see methods section 3.6). Specifically, we generated 100 minimal

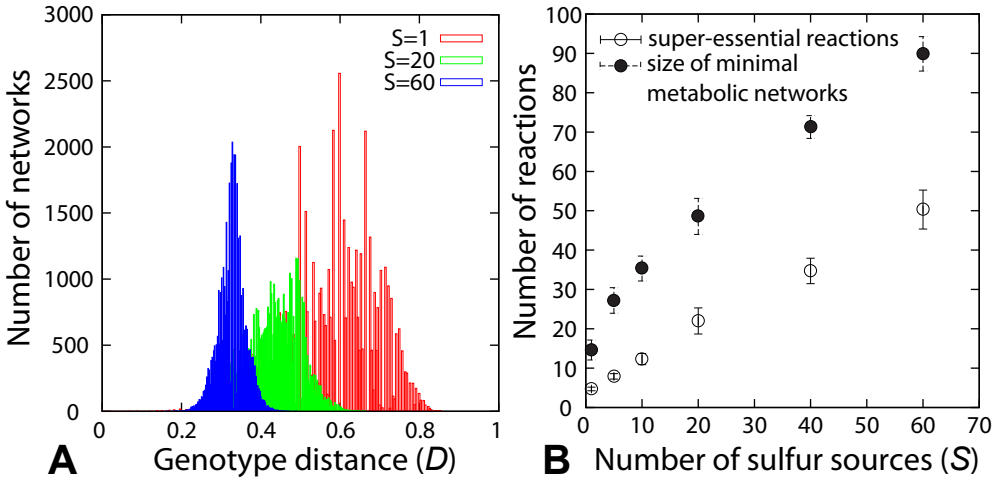

Figure 3.2: (A) Distribution of genotype distance between pairs of minimal metabolic networks viable under the same environmental demands. (B) Average size (closed circles) and average number of superessential reactions (open circles) of 1000 minimal metabolic networks as a function of environmental demands S on a network. The number of superessential reactions was obtained by counting the number of reactions common to 100 minimal metabolic networks generated with the same set of sulfur sources. For each data point, we used 10 different sets of sulfur sources with size S. Error bars represent the standard deviations of the distributions.

networks for 10 random sets of sulfur sources with the same number S — but not necessarily identity — of sources. We note in passing that such networks often also happen to be viable on additional sulfur sources that we did not require them to be viable on (Figure 3.6). Figure 3.2A shows the distribution of genotype distances for pairs of minimal metabolic networks viable on $S = 1, 20, 60$ sulfur sources. The figure demonstrates that, first, random minimal metabolic networks can be very different from one another. Their genotype distance may exceed $D = 0.8$, meaning that they may share fewer than 20 percent of reactions. Second, their average distance depends on the number of sulfur sources a network needs to be viable on. Specifically, the average genotype distance is largest $D_{avg} = 0.6$ for minimal metabolic networks viable on $S = 1$ sulfur source, and decreases to $D_{avg} = 0.3$ for networks viable on $S = 60$ sulfur sources. Third, the distribution of genotype distances is much wider for metabolic networks subject to few environmental demands ($S = 1$) where it ranges from $D_{avg} = 0.2$ to $D_{avg} = 0.8$, than for metabolic networks subject to many environmental demands ($S = 60$) where it ranges from $D_{avg} = 0.2$ to $D_{avg} = 0.4$. Figure 3.2B (filled circles) shows the average size of minimal networks N_{min} as a function of the number of sulfur sources they are viable on. It ranges from $N_{min} = 14$ reactions for $S = 1$ to $N_{min} = 87$ reactions for $S = 60$. By definition, all reactions in a minimal network are essential, but some of these reactions are special because they occur in all minimal networks viable on a given set of sulfur sources. We call these reactions *superessential* reactions [140]. The open circles in Figure 3.2B shows the number of superessential reactions as a function of the environmental demands S on a network. The number of superessential reactions R_{SE} increases with S, but it is generally much lower than the total number of reactions in a minimal metabolic network. For example, at $S = 1$, 4 out of 14 reactions are superessential. At $S = 60$, 44 out of 87 reactions are superessential. We will show that the number of superessential reactions plays an important role in one of our analyses below.

3.3.3 Many viable sulfur metabolic network genotypes are connected via paths that lead far through metabolic genotype space.

We next extended our previous work on carbon metabolism to ask about the existence of genotype networks in the space of sulfur-involving reactions, and of neutral paths that traverse such networks while preserving a metabolic phenotype. We define a neutral path as a series of mutations (reaction additions or deletions) that leave a phenotype intact (Figure 3.1B). We emphasize that we do not use the term neutrality in its meaning of unchanged fitness in the field of molecular evolution [144], but merely for brevity, in the sense of preserving viability on a

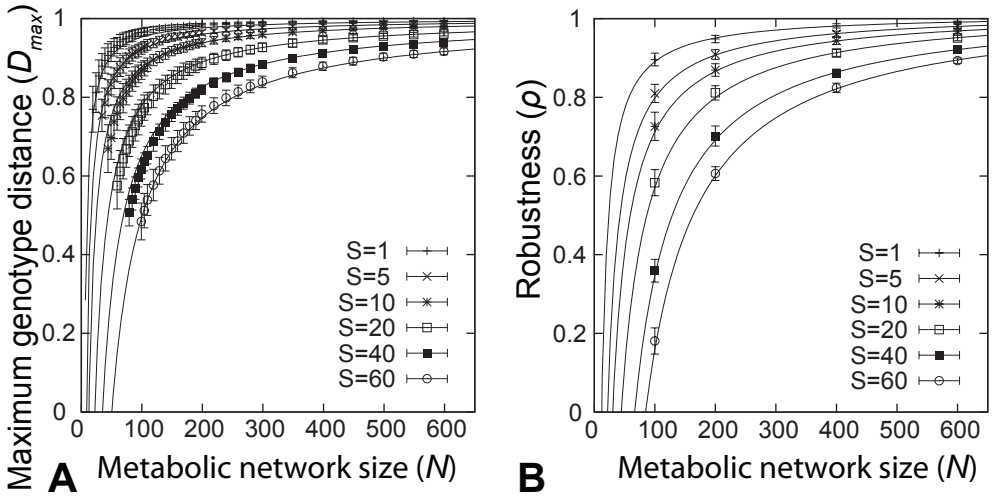

Figure 3.3: Maximum genotype distances and robustness depend strongly on the number of superessential reactions and the sizes of minimal metabolic networks. (A) Average maximum genotype distance for metabolic networks of different sizes and subject to different environmental demands after random walks of 10 000 accepted reaction changes. (B) Robustness of random metabolic networks with different sizes and subject to different environmental demands. Each data point is an average over 200 random walks (20 random walks for 10 different sets of environmental demands with the same number S of sulfur sources).

specific set of sulfur sources. Changes to metabolic networks such as additions or deletions of chemical reactions can potentially have a positive or negative effect on fitness. The addition of a chemical reaction may have a beneficial effect if it increases the rate at which biomass is synthesized, or it may have a deleterious effect if it generates metabolites that interfere with cell physiology. Similarly, the deletion of a reaction may have both beneficial and deleterious fitness effects [145]. Studies on compensatory mutations in macromolecules, mutations that compensate for the fitness effects of previous mutations with negative effects, show that fitness neutrality is not a prerequisite for a population's genetic change on a genotype network [146, 147]. We are especially interested in two questions. How far does a neutral path typically lead through genotype space? And how does this distance depend on the number N of reactions in a network, and on the environmental demands on the network? To answer these questions, we performed 200 random walks of 10 000 mutations each for metabolic networks of various sizes, and for various environmental demands. Specifically, for networks of each size we performed 20 random walks for each of 10 different sets of S of sole sulfur sources that we required the network to be viable on. Each random walk started from a random initial viable metabolic network comprising N reactions (see methods section 3.6 for details). We allowed N to vary by no more than one reaction during the random walk. Moreover, each step in the random walk had to preserve viability. Finally, none of the steps was allowed to decrease the distance to the starting network, in order to maximize the distance from this network (see methods section 3.6 for details). Figure 3.3A shows the maximum genotype distance D_{max} obtained in such random walks for networks up to 300 reactions, where we required viability on $S = 1, 5, 10, 20, 40, 60$ different sole sulfur sources. This distance is in general large. For example, D_{max} is greater than 0.7 for all metabolic networks with more than 200 reactions. For each value of S, the data point at the smallest value of N (horizontal axis) corresponds to the minimal metabolic networks we discussed earlier. Perhaps surprisingly, these minimal networks can not only be very diverse, as we saw earlier, but neutral paths starting from any one such network can also reach far through genotype space. For example, the maximal length of neutral paths is $D_{max} = 0.65$ for minimal metabolic networks viable on $S = 1$ sulfur source, and still a sizeable $D_{max} = 0.38$ for metabolic networks viable on $S = 60$ sulfur sources. To provide a point of reference, the *E. coli* metabolic network has 142 reactions involving sulfur. Random viable metabolic networks of this size would have maximum genotype distances between $D_{max} = 0.96$ (for $S = 1$) and $D_{max} = 0.60$ (for $S = 60$).

3.3.4 Maximal genotype distance and robustness of metabolic networks are well approximated by simple properties of minimal networks.

We asked whether the maximal genotype distance of networks of a given size, as well as their robustness to reaction removal can be predicted from properties of the underlying minimal networks. The answer is yes. Figure 3.3A shows that the maximal possible genotype distance D_{max} between metabolic networks of the same size increases with metabolic network size N. The solid lines show the relationship between the maximal genotype distance D_{max} and metabolic network size N as predicted by the equation

$$D_{max} = 1 - \frac{R_{SE}}{N}. \tag{3.2}$$

Here, R_{SE} is the number of reactions that are super-essential for a given environmental demand S. We had estimated this number in our previous analysis of minimal networks (Figure 3.2B). The simple relationship of equation (3.2) fits our numerical data (Figure 3.3A) remarkably well and corresponds exactly to our distance function when the number of superessential reactions R_{SE} replaces the number of common reactions R_C between networks of the same size. This implies that the only common reactions between maximally distant networks are the superessential reactions. Therefore, as the size of a network increases, more phenotype-preserving changes become possible. For networks of the smallest size, D_{max}, systematically overestimates the maximal genotype distance, but it does so by no more than 10% percent. We note that our estimates of maximum genotype distances are only lower bounds, such that this discrepancy may result from our limited ability to find maximal genotype distances accurately. In sum, a simple, linear function of the number of superessential reactions at any one environmental demand S approximates the maximal genotype distance between networks well. Next we examined how network robustness depends on the size of metabolic networks and on environmental demands. We define robustness as the fraction of non-essential reactions in a metabolic network. Figure 3.3B shows the robustness of metabolic networks as a function of network size N and varying environmental demands S on a network. Large metabolic networks with 200 reactions or more have a robustness $\rho > 0.6$ for all values of S. For smaller metabolic networks ($N < 200$), robustness ranges from $\rho = 0.9$ under low environmental demands ($S = 1$) to $\rho = 0.2$ under high environmental demands ($S = 60$). The relationship between ρ and network size N can be explained by noting that $\rho = 1 - R_{ess}/N$ where R_{ess} is the number of essential reactions. We find that R_{ess} decreases linearly with increasing metabolic network size (Figure 3.7) and is described by the

function $R_{ess} = N_{min}(1 + m) - Nm$, In this equation, N_{min} is the average size of minimal metabolic networks (estimated above for given S) and m is the rate at which the number of essential reactions decreases with increasing metabolic network size (estimated from data in Figure 3.7). The question why the number of essential reactions R_{ess} decreases with increasing size has a simple answer. As one increases the size of a metabolic network by adding reactions and entire pathways to minimal metabolic networks, some reactions may become non-essential because the added reactions create alternative pathways for biomass metabolite synthesis. Describing ρ in terms of N, N_{min} and m, we arrive at the following relation

$$\rho = 1 + m - \frac{N_{min}}{N}(1 + m)., \tag{3.3}$$

which is plotted as the solid lines in Figure 3.3B and fits the data very well. This relationship means that network robustness is a linear function of the ratio N_{min}/N, whose inverse indicates how much larger a given network is than a minimal network for a given S, and of the rate at which reaction essentiality declines (robustness increases) with increasing N.

3.3.5 The diversity of phenotypes found in the neighborhood of two metabolic networks changes rapidly with their genotype distance.

Thus far, we have concentrated on the characteristics of individual sets of genotypes viable on a given number S of sulfur sources, and on the genotype networks they form. Long paths through a genotype network can contribute to evolutionary innovation in metabolic phenotypes, if many novel phenotypes can be encountered near such a path. We next asked whether this is the case, and how this number of novel phenotypes depends on environmental demands on a network. We consider a phenotype to be novel if it confers viability in a set of new sulfur sources, in addition to those required by the environmental demands imposed on the metabolic network. We first introduce the notion of a (1-mutant) neighborhood around a metabolic network genotype, which comprises all networks that differ from the genotype by a single reaction (Figure 3.1B). Because our genotype space has 1221 metabolic reactions, each metabolic network has 1221 neighbors. Of all these neighbors, some will be inviable in any given environment (these are the mutants that have lost an essential reaction), some will maintain the same phenotype, and some will have a novel phenotype while being viable in this environment. That is, they will have gained viability on a new sulfur source. We focus on the latter class of neighbors in this section. We asked how different are the novel phenotypes in the neighborhood of two metabolic networks G and G_k on the same genotype net-

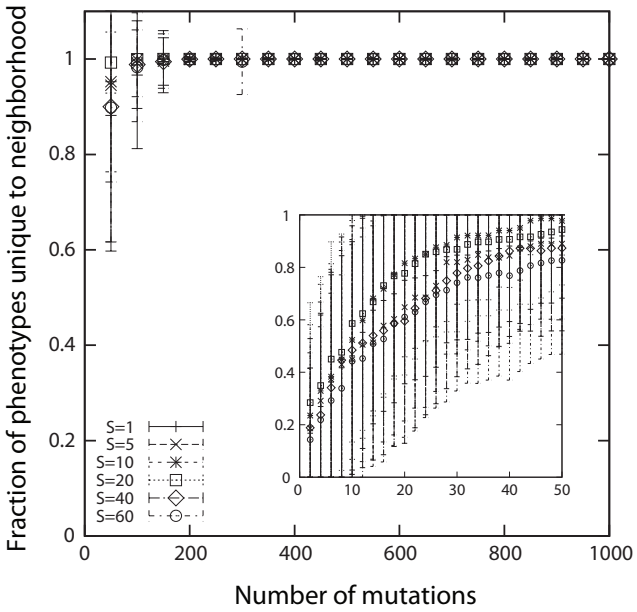

Figure 3.4: Fraction of phenotypes unique to the neighborhood of an evolving metabolic network G_k (vertical axis) when compared to the neighborhood of a starting metabolic network G after k "mutations", i.e., reaction changes (horizontal axis). Each data point represents an average over 200 evolving metabolic networks of size $N = 200$ (20 random walks for 10 different sets of environmental demands with the same number S of required sulfur sources).

work, where G_k is a metabolic network derived from G through k random reaction changes. That is, we determined the fraction of novel phenotypes that occurred in the neighborhood of only one but not the other network. Below we refer to it as the fraction of novel phenotypes unique to one neighborhood. If this fraction is very small even for large k, then networks in different regions of a genotype space will have mutational access to similar novel phenotypes. Figure 3.4 shows that the opposite is the case. We obtained the data shown during phenotype-preserving random walks starting from an initial network, by recording the fraction of novel phenotypes that occur in the neighborhood of the changing metabolic network, but not of the initial network. Every data point is an average over 20 random walks each for 10 different initial metabolic networks (thus, 200 random walks in total) at every value of S. Figure 3.4 shows that the fraction of unique novel phenotypes reaches high values for modest distance between two metabolic networks — small compared to the maximum genotype distance — and does not depend much on the number of sulfur sources S on which viability is required. It also does not depend strongly on metabolic network size (results not shown). In sum, the neighborhood of moderately different metabolic networks contains very different novel phenotypes.

3.3.6 The ability of metabolic networks to encounter novel phenotypes does not depend monotonically on their phenotypic robustness.

The question of how robustness relates to phenotypic variability has raised considerable interest in recent years [148, 75]. Macromolecules — RNA and protein — whose phenotypes are more robust to mutations can access more novel phenotypes than less robust phenotypes [141, 74]. This holds for both large and small evolving populations of such molecules, at least in RNA [74]. We next asked whether these observations are specific to macromolecules, or whether they hold more generally, that is, also for the genotype-phenotype map of metabolic networks. Above we considered the robustness of a metabolic genotype as its fraction of non-essential reactions. Analogously, we now consider the robustness of a metabolic phenotype as the average fraction of non-essential reactions of all networks with this phenotype [74]. We showed that robustness decreases as networks are required to be viable on more and more sulfur sources (Figure 3.3B). That is, for networks of any given size, the number S of sulfur sources on which they are viable can serve as a proxy for phenotypic robustness. The greater a phenotype's S is, the smaller is its robustness. When analyzing how evolving populations explore a genotype network, we need to distinguish between two different kinds of populations. The first are populations where the product of population size and mutation rate is

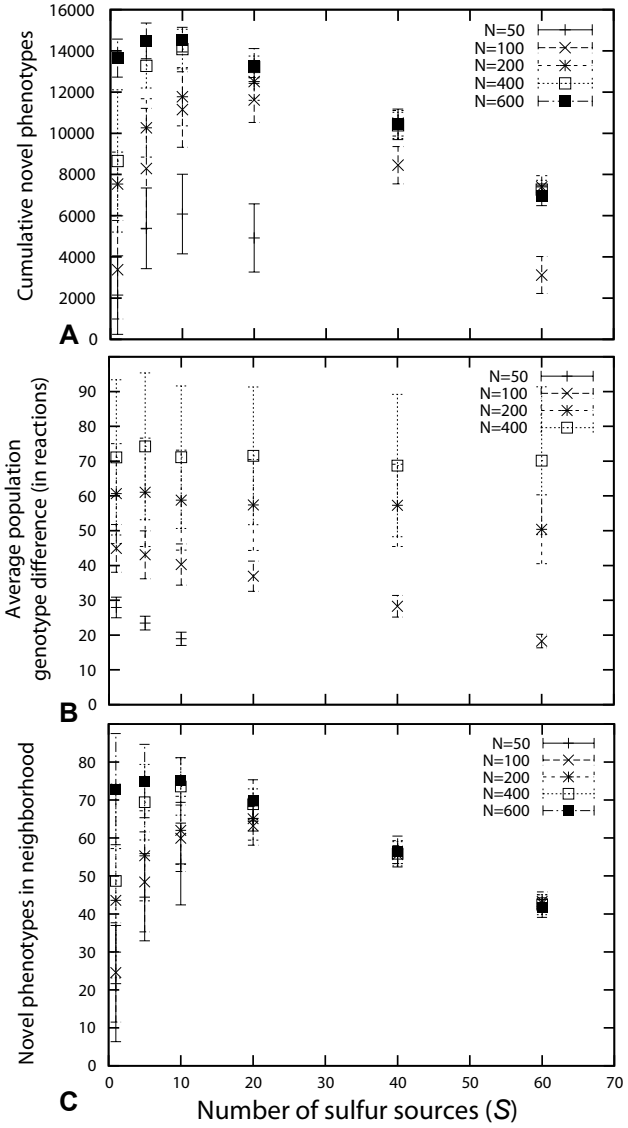

Figure 3.5: Metabolic networks have diverse phenotypes in their neighborhood. (A) Cumulative number of novel phenotypes encountered in the neighborhoods of evolving metabolic networks of different sizes and subject to different environmental demands. (B) The average pairwise genotype distance found in populations of evolving metabolic networks. Each population consists of 100 individual metabolic networks. (C) Number of novel phenotypes found in the (1-mutant) neighborhood of random metabolic networks of different size N and subject to different environmental demands S. Each data point is an average over 200 metabolic networks (20 random walks for 10 different sets of environmental demands, with the same number S of required sulfur sources).

much smaller than one. For brevity, we refer to such populations here as small populations. The second are populations where this product is much greater than one. We refer to these as large populations. Small populations are genotypically monomorphic most of the time [144], and effectively explore a genotype network much like a single changing network would, i.e., through a random walk on the genotype network. During such a random walk, the changing network encounters different phenotypes in its neighborhood. We determined the cumulative number of different novel phenotypes found in the neighborhood of a random walker. That is, if a phenotype was encountered twice, either in the same neighborhood, or in a neighborhood encountered during an earlier step, we counted it only once. We did so for networks of varying size N and number of sulfur sources S. Specifically, for each N and S, we carried out 200 random walks of 10 000 mutations each (20 walks for 10 different sets of sulfur sources at each S). Figure 3.5A shows the resulting data. The cumulative number of novel phenotypes is a unimodal function of S, indicating that metabolic networks under few and many environmental demands encounter fewer novel phenotypes than under an intermediate number of environmental demands ($S \approx 20$). The cumulative number of novel phenotypes depends strongly on metabolic network size for $S < 20$, where larger metabolic networks encounter more novel phenotypes throughout the random walk. It is not sensitive to N for larger values of S. We next examine large evolving populations. Such populations are polymorphic most of the time. To model their evolutionary dynamics, one needs to track every individual in the population, unlike for monomorphic populations. We determined the cumulative number of novel phenotypes that are mutationally accessible to a population of metabolic networks evolving on (and restricted to) a specific genotype network. This number can be determined by examining, for each generation, the neighborhood of each individual in the population, and by counting the total number of different novel phenotypes encountered. We simulated populations of 100 individuals evolving for 2000 generations (see methods section 3.6 for details). Figure 3.8 shows the average number of cumulative unique novel phenotypes accessible to a population through generation 2000. Each data point represents an average and standard deviation over 200 simulations (20 simulations for 10 random sets of sulfur sources at a given S). Qualitatively, the figure resembles our observations for a single random walk (Figure 3.5A), except that the absolute number of cumulative unique phenotypes encountered is higher in evolving populations. Taken together, these observations show that the number of novel phenotypes accessible to a population does not increase monotonically with phenotypic robustness. It decreases with increasing robustness (decreasing S) for low values of S, and it increases with robustness at higher values of S. Figure 3.9 demonstrates this relationship in a 3D plot of the number of novel phenotypes versus robustness and network size (A) or environmen-

tal demands (B). We next examined two candidate explanations of this pattern. The first is that environmental demands and network size affect how rapidly a population can diversify on its genotype network, and thus also how many novel phenotypes it can access. To find out whether this diversification rate matters, we examined the average pairwise genotype distance of our evolving populations. The smaller this difference is, the more slowly a population diversifies. Figure 3.5B shows a plot of pairwise genotype distances, averaged over an entire population, at the end of 2000 generations. One can see that populations of smaller networks are less diverse. However, environmental demand (S) influences genotypic diversity only weakly, and not in the same unimodal way as seen in Figure 3.8. Thus, population dynamic processes alone cannot explain the pattern observed in Figures 3.5A and 3.8. The second candidate explanation is that the patterns of Figures 3.5A and 3.8 may simply reflect how the number of novel phenotypes in the neighborhood of random metabolic networks varies with N and S. Figure 3.5C shows the number of novel phenotypes in the neighborhood of random viable metabolic networks of varying size, and with varying environmental demands on the network. This figure is based on random samples of 200 metabolic networks (see methods section 3.6) for every value of N and S (20 metabolic networks for 10 different sets of sulfur sources at each S). The vertical axis of this figure shows the mean and standard deviation of the number of unique novel phenotypes in the neighborhood of the examined networks. It shows similar unimodal characteristics as the data in Figures 3.5A and 3.8. The figure demonstrates that the number of novel phenotypes depends strongly on metabolic network size for environments with $S < 20$. In this regime, larger metabolic networks have more novel phenotypes in their neighborhood than smaller networks. For $S > 20$, the dependency on metabolic network size disappears and the number of accessible novel phenotypes declines again. In sum, the different accessibility of novel phenotypes in evolving populations, at least qualitatively, emerges from how novel phenotypes are distributed in genotype neighborhoods, and how this distribution depends on S and N.

3.4 Discussion

The genotype-phenotype map we characterized here shows both similarities and differences to previously characterized such maps [61, 62, 63, 139, 140]. One similarity is the existence of connected genotype networks that extend far through genotype space, and that link genotypes having the same phenotype. Connected sets of metabolic networks viable on the same set of sulfur sources exhibit large maximum genotype distances D_{max}. For example, networks with as few as 200 reactions can show $D_{max} > 0.7$, meaning that they share fewer than 30 percent of their reactions. A second similarity regards phenotypic innovations, genotypes

whose phenotypes allow viability on novel sulfur sources. The neighborhoods of two genotypes G_1 and G_2 tend to contain very different phenotypic innovations, even if G_1 and G_2 are only moderately different. Both features, taken together, facilitate the exploration of novel phenotypes. They would allow a population of organisms (networks) to explore different regions of genotype space, preserving their phenotype while exploring many novel phenotypes. A major difference to previously studied genotype-phenotype maps regards the relationship between a phenotype's robustness to mutation and a population's ability to explore novel phenotypes. In macromolecules, this relationship appears to be positive: Greater robustness facilitates innovation [141, 74]. Although robust molecules can access, on average, fewer novel phenotypes in their mutational neighborhoods, populations of robust molecules can spread faster through genotype space. In balance, the second process dominates and allows evolving populations to access more novel phenotypes through mutations. In sulfur metabolism, we do not see this relationship. Robust phenotypes in this context are characterized by viability on few sulfur sources. They are less easily disrupted through eliminations of individual reactions. We found that the number of phenotypic innovations such phenotypes can access in their neighborhood — through changes of single reactions — is highest at intermediate robustness, that is, for phenotypes viable on approximately 20 out of 60 carbon sources we examined. (It can also depend on metabolic network size, being lowest for small networks.) This phenomenon cannot solely be explained by the evolutionary dynamics of evolving populations, partly because populations whose members have intermediate robustness do not spread fastest through genotype space. Instead, the phenomenon is a simple consequence of how many novel phenotypes occur in the neighborhoods of individual genotypes. This number peaks for genotypes whose phenotypes have intermediate robustness. It shows the same qualitative dependence on robustness as the number of novel phenotypes accessible to populations. Thus, in this case, population dynamics do not dominate the process of novel phenotype exploration. We note that the total number of possible novel phenotypes decreases exponentially with the number S of sulfur sources on which a network is already viable. If we took this exponential decrease into account, for example by determining the cumulative fraction instead of the number of novel phenotypes accessible to evolving population, this fraction would decrease with increasing S. These observations raise the question whether they are unique to sulfur metabolism or whether they occur in other metabolic systems. As we stated earlier, part of our motivation to study sulfur metabolism was to avoid the much larger number of reactions of carbon metabolism, which render population approaches like ours computationally intractable. Nonetheless, very limited analyses for carbon metabolism are possible. Figures 3.10A and 3.10B show the results of such an analysis, based on a small number of populations of

networks at moderate size. The analysis has large uncertainties, but it shows a pattern that is at least reminiscent of sulfur metabolism: Innovation peaks at intermediate robustness (the number of alternative carbon sources a phenotype is viable on). Taken together these analyses show that the organization of different phenotypes in genotype space can differ greatly among different classes of biological systems, such as proteins and metabolic networks. And these differences can affect the ability of a system to explore novel phenotypes in this genotype space. A third class of analyses regards features that have not been studied previously, partly because they are unique to metabolic systems and our representation of them. One of them regards the analysis of networks with different sizes (numbers of reactions). Our genotype representation can accommodate and allows us to compare networks of different sizes, whereas commonly used representations of other systems — molecules or regulatory circuits — cannot. For example, proteins of different length form genotype spaces of different dimensions, making their comparison challenging [149]. When analyzing metabolic networks of different sizes, we found that populations of small networks can explore fewer novel phenotypes (Figures 3.5A and 3.8). This observation is easily explained if one considers that populations of such networks are less robust. Their genotype can thus be altered less easily. In consequence, they are genotypically less diverse (Figure 3.5B), which restricts their access to novel phenotypes (Figure 3.5B). Another analysis focusing on network sizes is our characterization of minimal metabolic networks, networks in which all reactions are essential. While the process of genome and metabolic network reduction leading to small networks has been studied for specific biological networks [106], our approach does not start from such a network and can thus provides a more systematic exploration of genotype space. In our analysis of random minimal metabolic network viable on the same sulfur sources, we found that such networks can have large genotype distance. We can explain part of this observation through reactions that are very similar but differ in one of several highly related metabolites. For example, in many types of reactions involving the phosphorylation of a metabolite, the phosphor group donor can be any of ATP, ADP, AMP or even other phosphorylated nucleotide bases. This allows single reactions to be substituted by similar reactions that only use another group donor metabolite. Also, many alternate pathways require only the swapping of two reactions allowing metabolic networks with very little robustness to substitute some of their reactions. However, these may not be the only explanations of different network architectures, because minimal metabolic networks viable on the same sulfur sources can have dramatic pathway differences (results not shown). Whether such differences can be bridged through series of single reaction changes is a question for future exploration. Properties of minimal networks are also useful in explaining the maximal genotype distance in a genotype network. For example,

for metabolic networks of a given size N and viability on S sulfur sources, the maximum genotype distance within a genotype network is well approximated by one minus the fraction of superessential reactions in minimal metabolic networks. These are reactions found in all minimal networks viable on a given number of sulfur sources. We currently have no mechanistic explanation for this relationship and it, also, remains a subject for future work. Studies like ours have several limitations. One of them is that we focus on biomass synthesis phenotypes, and not on other aspects of metabolism, such as secondary metabolite production. The reason is that biomass synthesis has the most immediate impact on an organism's survival. Other limitations include that the addition and deletion of reactions may have effects on fitness even if they do not affect biomass synthesis, that our knowledge of the reaction universe is limited, and that we face uncertainty about the biologically most important sulfur sources, about thermodynamic properties of individual metabolic reactions, and about the role of cellular compartmentalization in guiding sulfur metabolism. Our study, even though it uncovers generic features of genotype-phenotype maps with demonstrated relevance for evolutionary adaptation and innovation in other biological systems [141, 150], is thus best viewed as a modest beginning in characterizing a complex metabolic genotype space.

3.5 Conclusions

We demonstrate that metabolic networks in sulfur metabolism with the same phenotype form large genotype networks that reach far through metabolic genotype space. How far they reach through this space is a linear function of the number of super-essential reactions specific to the environment. We show that the robustness of metabolic networks depends on the size of a metabolic network, on the average size of minimal networks viable in a given environment, and on how rapidly the proportion of essential reactions decrease with increasing network size. The neighborhoods of two metabolic networks on the same genotype network typically contain different novel phenotypes. In evolving populations of metabolic networks, robustness facilitates the discovery of novel phenotypes only up to some modest value of robustness, beyond which populations discover fewer novel phenotypes. The difference in the role of robustness in the evolution of metabolic networks compared to its role in the evolution of macromolecules shows that phenotypic innovations may not occur according to the same principles in all biological systems.

3.6 Methods

3.6.1 Global set of sulfur-involving reactions

To obtain the global set of reactions involving sulfur-containing metabolites that can be present in the metabolic networks we studied, we used data from the LIGAND database of the Kyoto Encyclopedia of Genes and Genomes (KEGG; http://www.genome.ad.jp/kegg/ligand.html) [50]. The LIGAND database is a database of chemical compounds and reactions in biological pathways that was compiled from pathway maps of metabolism of carbohydrates, energy, lipids, nucleotides, amino acids and others. Also included in the database are the list of catalyzed reactions categorized by the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) which includes all enzymes with known classification [151]. Specifically, we used the REACTION and COMPOUND sections of the LIGAND database to construct our global reaction set. From this dataset we pruned (i) all reactions involving general polymer metabolites of unspecified numbers of monomer units ($C_2H_6(CH_2)_n$), or, similarly, general polymerization reactions that were of the form $A_n + B \rightarrow A_{n+1}$, because their abstract form makes them unsuitable for stoichiometric analysis, (ii) reactions involving glycans, because of their complex structure, (iii) reactions that were not stoichiometrically or elementally balanced, and (v) reactions involving complex metabolites without chemical information about their structure. In addition, we merged all the reactions existing in the E. coli metabolic network model (iJR904) [95] that involve sulfur containing compounds. After these steps of pruning and merging, our global reaction set consisted of 1221 reactions.

3.6.2 Flux balance analysis

Flux balance analysis is a computational method used to find a set of fluxes through all metabolic reactions that maximize biomass production in a given metabolic network, assuming it is in a steady state [46]. This assumption means that the concentrations of internal metabolites does not change over time. To compute the maximum biomass growth using this method, one needs to know the stoichiometric coefficients of each reaction, the chemical environment of the cell (the set of upper bounds on the fluxes of external metabolites into the cell), and the biomass composition, which represents metabolite consumption during cell growth. This consumption is reflected in a "biomass growth reaction", for which we chose the reaction defined in the E. coli iJR904 metabolic model [95]. This biomass growth reaction includes all 20 proteinaceous amino acids, nucleotides, deoxynucleotides, putrescine, spermidine, 5-methyltetrahydrofolate, coenzyme-A, acetyl-CoA, succinyl-CoA, cardiolipin, FAD, NAD, NADH, NADP, NADPH, glycogen,

lipopolysaccharide, phosphatidylethanolamine, peptidoglycan, phosphatidylglycerol, phosphatidylserine and UDPglucose. For the purpose of this study we concentrated only on the ability of a metabolic network to synthesize the sulfur containing biomass precursors, which are the two amino-acids cysteine and methionine, coenzyme-A, acetyl-CoA and succinyl-CoA. We thus allowed the metabolic networks to uptake any metabolites not containing sulfur. We consider a metabolic network to be viable in a given environment if it can sustain a biomass growth rate greater than 1.0×10^{-3}. In essence, the approach we take is equivalent to asking whether all the necessary sulfur containing biomass precursors are synthesizable given a metabolic network in a specified environment. Flux balance analysis relies on linear programming [55] to compute the maximum biomass production rate. We used the packages CPLEX (11.0, ILOG; http://www.ilog.com/) and CLP (1.4, Coin-OR; https://projects.coin-or.org/Clp) to solve the associated linear programming problems.

3.6.3 Environments and phenotypes

We here considered 124 different environments that differed in the chemical compound that could serve as the *sole* source of sulfur. These 124 sources were all the sulfur containing metabolites in the 1221 reactions of our global reaction set. We provided any metabolite not containing sulfur in the environment, in effect making it a rich environment limited by sulfur containing metabolites only. Also, we allowed cells to secrete all metabolites. We define a metabolic *phenotype* as the set of environments (each with a different sole sulfur source) in which a metabolic network is viable. The *environmental demands* imposed on a metabolic network correspond to the set of sulfur sources that the metabolic network must *at least* be viable in.

3.6.4 Essential and super-essential reactions

We define a reaction as *essential* if its removal from a metabolic network renders the metabolic network inviable on at least one of the sulfur sources that it had previously been viable on. We called a reaction *super-essential* if it occurred in all minimal metabolic networks generated under a given set of environmental demands.

3.6.5 Generating random and minimal metabolic networks

We generated random viable metabolic networks as follows. First, we generated a random environmental demand, that is, we required viability in some given number X of sulfur sources. To this end, we first created a binary vector of length 124

(each of whose entries corresponds to one sulfur source), initialized all its entries to the value zero, and then randomly changed X of these entries to one. These entries represent the set of sulfur sources on which we required our metabolic networks to be viable.
We then generated random viable metabolic network of N reactions as follows. We started from a metabolic network that contained all 1221 reactions (this networks is viable on all 124 sulfur sources) and sequentially removed randomly chosen reactions, while ensuring viability on the set of X sulfur sources chosen previously, until we had reached a network with the target number N of reactions.
We define a *minimal metabolic network* as a network where not a single reaction can be removed without destroying viability. To generate a (random) minimal metabolic network we used the same procedure until no reactions could be removed without destroying viability.

3.6.6 Metabolic network random walk maintaining viability in the environmental demands

We generated random walks for metabolic networks of given reaction numbers N and viability on a given number of sulfur sources by first generating a random metabolic network of this size, as just described. We then generated a series of steps ("mutations") in metabolic genotype space, each one either an addition or a deletion of a reaction. After each step, we recomputed the phenotype of the metabolic network. If the metabolic network was still viable on the same set of sulfur sources, we accepted the mutation and proceeded to the next mutation; if not, we rejected the mutation and repeated the process from the metabolic network prior to the mutation. We continued the resulting random walk for 10 000 accepted mutations. We kept the size of the metabolic network in the narrow interval $(N, N+1)$ by ensuring that accepted mutations alternated between reaction additions and deletions. In a variation on this procedure, we also carried out forced random walks through genotype space. Their aim was to obtain metabolic networks that are as different (in terms of genotype distance) as possible from the initial metabolic network. In a forced random walk, we required that any reaction addition did not involve a reaction that had been part of the initial network at the start of the walk.

3.6.7 Population dynamics

Populations where the product of population size and mutation rate is much greater than one are polymorphic most of the time, and show evolutionary dynamics different from those of small populations [12]. To understand their evolution,

one needs to simulate them explicitly. To this end, we implemented a Fisher-Wright model of evolution [152] in populations of 100 metabolic networks. We initialized each population with 100 copies of a single viable metabolic network, and then exposed it to repeated "generations" of mutation (one reaction addition or deletion per network and generation) and selection. Specifically, for the selection procedure, we chose 100 viable individuals at random with replacement to form the next generation. If a mutation had rendered a network inviable, it could not be chosen. Our simulations proceeded for 2000 generations.

3.7 Supplementary Figures

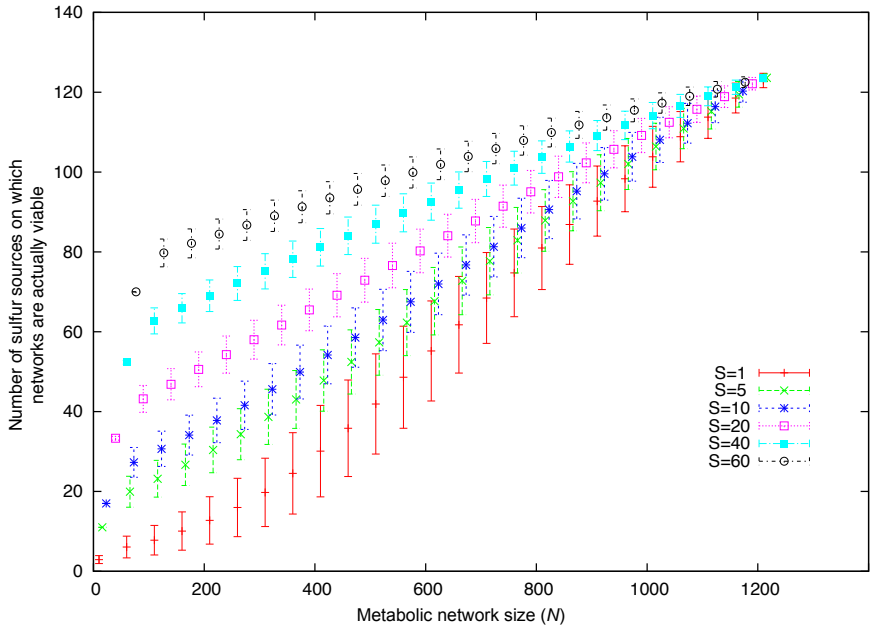

Figure 3.6: Average number of sulfur sources that random metabolic networks are actually viable in, for varying environmental demands S, and varying metabolic network size N. The figure demonstrates that random metabolic networks required to be viable on a given number S of sulfur sources (as generated by the procedures described in Methods) are generally viable on more than S sulfur sources. Each data point represents an average over 200 random metabolic networks (20 random metabolic networks generated under 10 different sets of environmental demands with the same number S of required sulfur sources). Error bars correspond to one standard deviation.

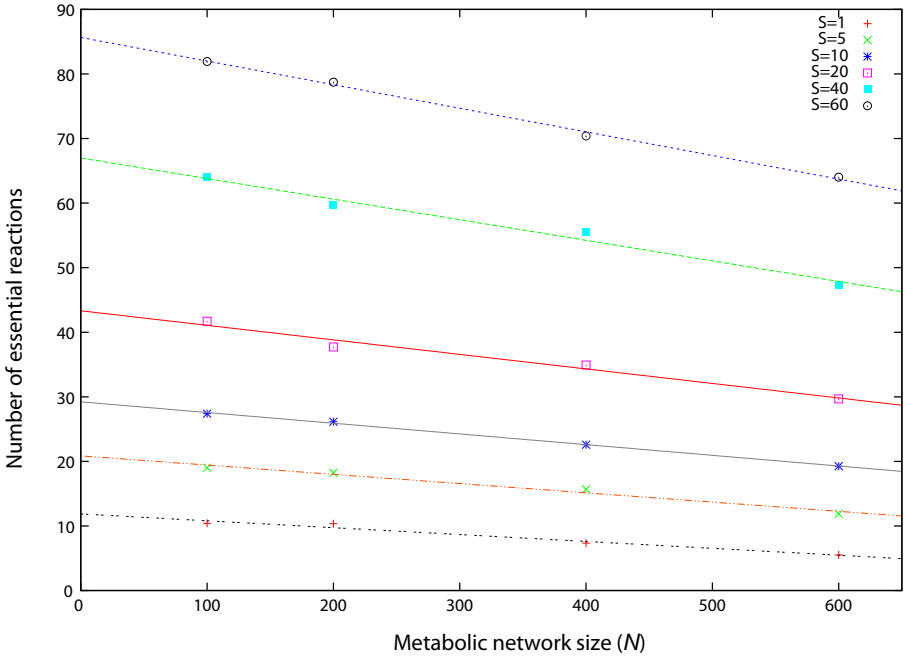

Figure 3.7: Number of essential reactions found in random metabolic networks of different size and for different environmental demands (S). Each data point represents an average over 200 random metabolic networks (20 random metabolic networks generated under 10 different sets of environmental demands with the same number S of required sulfur sources).

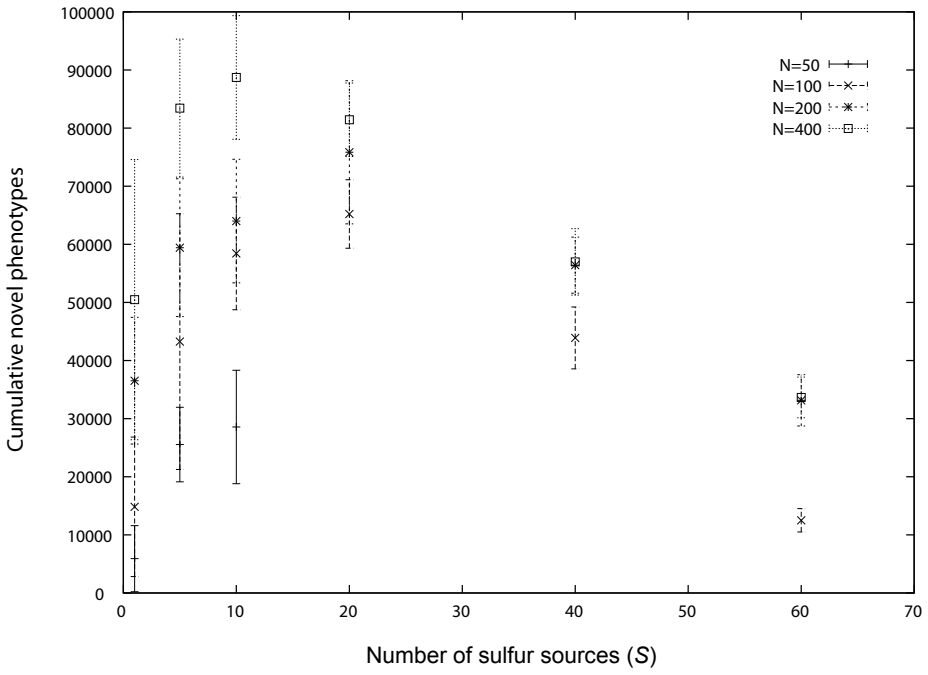

Figure 3.8: Cumulative number of novel phenotypes encountered in the neighborhoods of all evolving metabolic networks in a large population. The results are plotted for populations of metabolic networks of different sizes and subject to different environmental demands. Each data point represents an average over 200 simulations, 20 simulations for 10 different sets of environmental demands with the same number S of sulfur sources. Each population consisted of 100 individual metabolic networks.

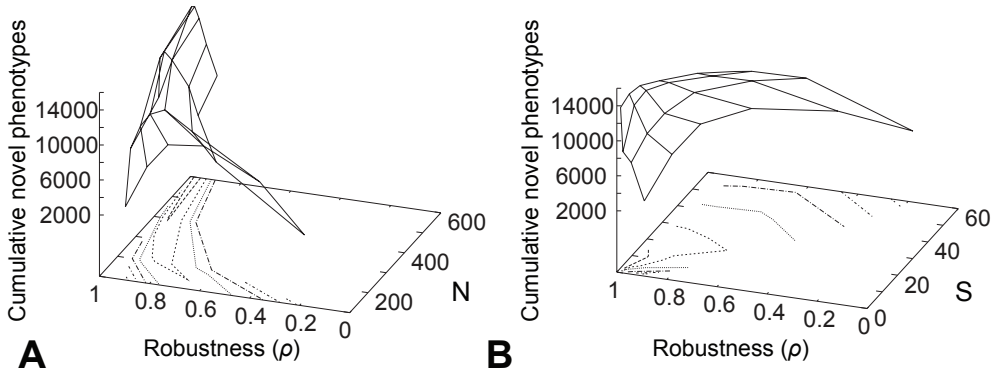

Figure 3.9: Three dimension plots of number of novel phenotypes found in the (1-mutant) neighborhood of random metabolic networks versus robustness of the metabolic networks and (A) metabolic size N or (B) different environmental demands S. Each data point is an average over 200 metabolic networks (20 random walks for 10 different sets of environmental demands, with the same number S of required sulfur sources).

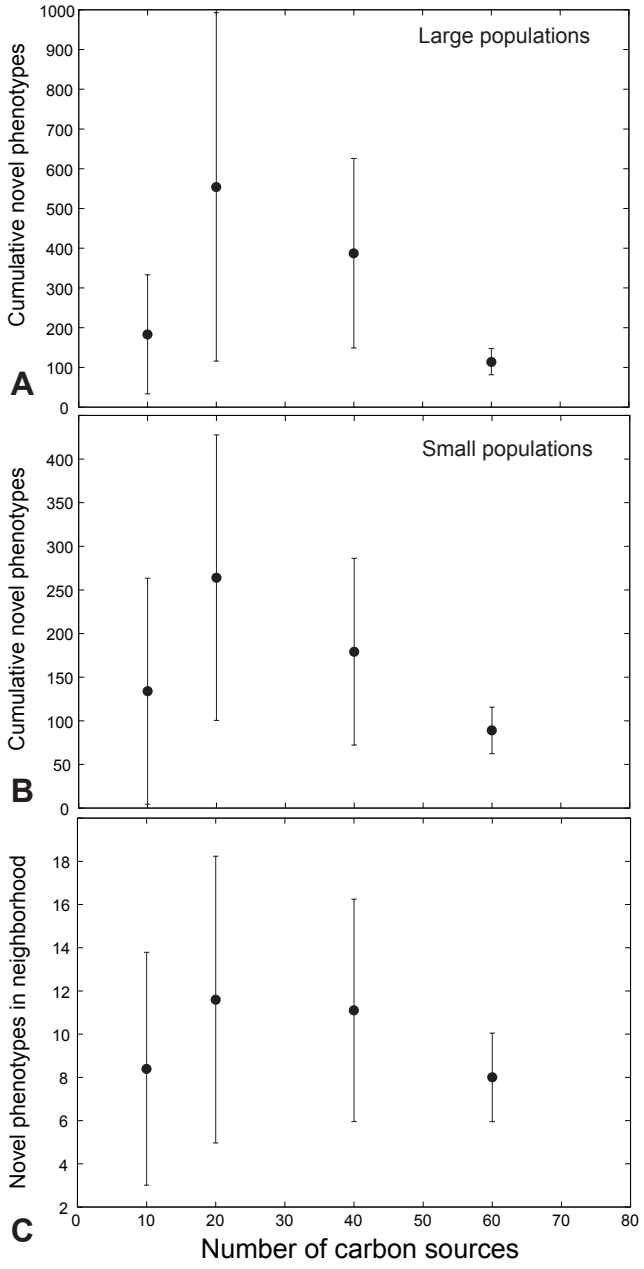

Figure 3.10: Plot of the cumulative number of novel phenotypes found in the neighborhood of (A) large and (B) small evolving populations of metabolic networks required to be viable in different number of carbon sources. (C) Number of novel carbon utilization phenotypes found in the neighborhood of random metabolic networks. Metabolic networks in these simulations had 931 reactions, the same as the size of the *E. coli* iJR904 model [95, 139].

4 Differential cell growth drives the evolution of terminal and reversible differentiation

João F. Matias Rodrigues, Daniel J. Rankin, Valentina Rossetti, Andreas Wagner, and Homayoun C. Bagheri
[*submitted*, **2010**]

4.1 Abstract

Multicellular differentiated organisms are composed of different cell types that develop from a single pluripotent germ cell. In many organisms, a proportion of cells differentiated into somatic cells. Whether these cells maintain their pluripotency and are able to reverse their differentiated state has important consequences. Reversibly differentiated cells can potentially regenerate parts of an organism and allow reproduction through fragmentation. In many organisms however, differentiation is terminal, thereby restricting the developmental paths to reproduction. The reason why terminal differentiation is a common developmental strategy remains unexplored. To understand the conditions that affect the evolution of terminal versus reversible differentiation, we develop a computational model inspired by differentiating cyanobacteria. We simulate the evolution of a population of two cell types —nitrogen fixing or photosynthetic— that exchange resources. The traits that control differentiation rates between cell types are allowed to evolve in the model. We find that although the topology of cell interactions and differentiation costs can play a role in the evolution of terminal and reversible differentiation, the most important factor is the difference in growth rates between cell types. Specifically, faster growing cells always become the germ line. Our results provide insights as to why some multicellular differentiated cyanobacteria are composed of reversibly differentiated cells, while other cyanobacteria have terminally differentiated cells. We further observe that symbioses involving two cooperating lineages can evolve under conditions where aggregate size, connectivity, and differentiation costs are high. This may explain why plants engage in symbiotic interactions with diazotrophic bacteria. The results are robust with regard to the type of resource exchange considered and can apply to a range of other systems.

4.2 Introduction

The reproduction and development of differentiated multicellular organisms follows a complex iterative pattern. Almost all differentiated multicellular organisms develop from a single pluripotent germ cell that divides and differentiates. The ability of differentiated cell-types to produce other cells through reversible differentiation determines an organism's mode of reproduction. Many organisms are composed of both germ cells and terminally differentiated somatic cells. The latter lose their ability to differentiate into other cell types. Although terminally differentiated somatic cells contain all the necessary genetic information to produce whole organisms [153, 154, 155], they are unable to do so despite the potential cost in reproductive opportunities for the organism. In contrast, organisms composed of reversibly differentiated cells can reproduce through fragmentation or budding.

Examples include most plants, and some animals such as corals, *hydra*, planarians, several echinoderms, and some annelid worms able to reproduce by fragmentation [84, 82, 83, 156]. In these organisms, each fragment regenerates the missing parts of the organism, resulting in several complete new organisms. During such regeneration, somatic cells in the fragments can sometimes de-differentiate and form a blastema (a group of undifferentiated cells) that regenerates the missing parts [83]. This means that somatic cells undergo reversible differentiation, because they are able to revert back to their undifferentiated forms.

Multicellular cyanobacteria are some of the simplest multicellular organisms known. They are of particular interest because in some species, cells are terminally differentiated [88], while in others, terminally differentiated cells have not been observed. Cyanobacterial species exist in many different morphologies. They are found as single cells, multicellular filaments of undifferentiated cells, and differentiated multicellular filaments (with or without branching) [91]. In differentiated multicellular cyanobacteria, some cells specialise in photosynthesis while others specialise in nitrogen fixation. Only one genus of cyanobacteria that could potentially exhibit reversible differentiation (*Trichodesmium*) is known [89, 92]. In contrast, several terminally differentiating cyanobacteria are known, of which two examples are the genera *Anabaena* and *Nostoc*. These cyanobacteria are composed of two cell types: the vegetative cell (germline) and the heterocyst cell (somatic). The vegetative cell is photosynthetic, reproduces through division and is able to differentiate into heterocyst cells [93]. The heterocyst cell is larger than the vegetative cell, has a thicker cell wall composed of three layers, and performs nitrogen fixation. In this manner, vegetative cells obtain fixed nitrogen from heterocysts, and heterocysts obtain fixed carbon from the vegetative cells. It has been suggested that the reason for heterocysts to be terminally differentiated is a consequence of their thicker cell membrane. However, the existence of a species of cyanobacteria such as *Trichodesmium* where cells perform nitrogen fixation and are capable of cell division [89, 92], suggests the possibility of other explanations.

The reasons why multicellularity or cell differentiation have evolved have received some attention recently [77, 78, 79, 81, 157, 158], but the evolutionary forces that drive the evolution of terminal versus reversible differentiation remain unexplored. Terminal differentiation can be seen as a case in which the differentiated individual pays a cost (in terms of lost reproduction) in order to confer a benefit on other cells nearby. This could be regarded as altruism *sensu* Hamilton [159]. Because selection acts at the level of the individual, altruism is vulnerable to cheating when interacting individuals are not closely related. However, as will be elaborated in the discussion section, some of the expectations based on such an interpretation do not match our results.

Using a spatially explicit approach we model the evolution of differentiation.

Our model follows assumptions about multicellular cyanobacterial species, but is nonetheless sufficiently general to apply to other systems. We assume that the physiological interaction of cells with neighbouring cells affects their reproductive success. We find that the topology of interactions, the differentiation costs, and the relative growth rate between different cell types can all play a role in the evolution of terminal or reversible differentiation. In addition, we find that some conditions can lead to the "speciation" of a multicellular organism into a symbiotic pair. In this case, the different cell types separate into two lineages evolving independently from each other. Our approach helps to identify some of the principal factors that led to the evolution of the diverse differentiation strategies seen in cyanobacteria.

4.3 Model

For this model we draw inspiration from the exchange of resources between cells in differentiated cyanobacteria. We consider a finite population of individuals or cells arranged in linear chains or filaments that exchange carbohydrates and fixed nitrogen in each iteration with some of their neighbours (Figure 4.1). After each round of interactions, the fitness of each individual is computed. Then the evolution of the population proceeds in a series of iterations composed of two steps. First, an individual is randomly selected for reproduction with a probability proportional to its fitness. Second, another individual is selected randomly for cell death, irrespective of its fitness.

An individual can be one of two cell types, either a photosynthetic cell or a nitrogen fixing cell. Each cell type produces only one type of resource (carbohydrates or fixed nitrogen). Because cells are composed of both carbon and nitrogen, they need both elements in order to grow and divide. Since the cell composition ratio of carbon to nitrogen (C:N) has been estimated to be around 6:1 for bacterioplankton [160] and a typical molecule of sugar produced in photosynthesis contains 6 carbon atoms, we consider the biomass composition to be 1 unit of carbohydrates to 1 unit of fixed nitrogen. Assuming that this ratio of biomass remains constant in the cell, and that cells require carbohydrates and fixed nitrogen in equal parts, their growth rate will therefore be limited by the least available resource. Since nitrogen fixation requires a large amount of energy we assume that nitrogen fixing cells require carbohydrates to perform nitrogen fixation. This stands in contrast to photosynthetic cells which are able to produce carbohydrates irrespective of the amount fixed nitrogen received.

To investigate the effect of differences in growth rates between the two cell types we define the parameter α which expresses how much faster a photosynthetic cell grows relative to a nitrogen fixing cell. Differences in cell growth rate may result from differences in cell composition or cellular structures or from other factors

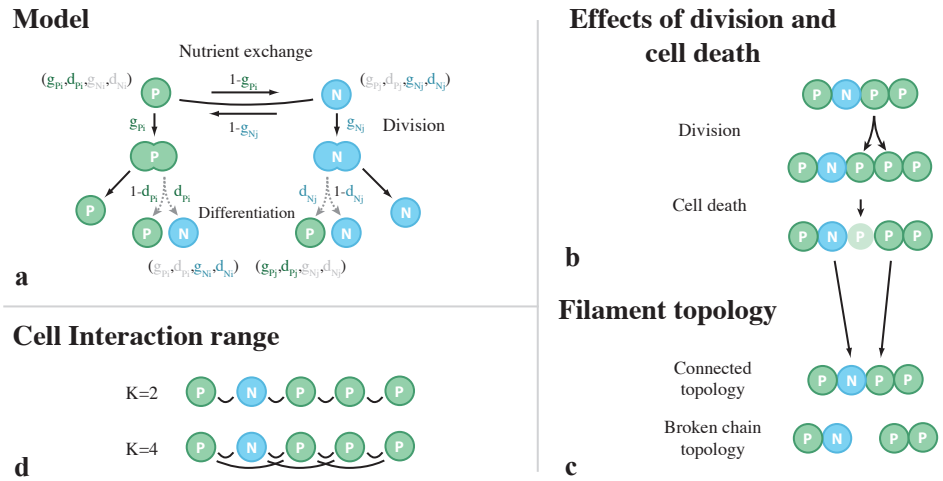

Figure 4.1: (a) Cells can be of two types, either photosynthetic (P) or nitrogen fixing (N). Every cell has 4 traits g_P, d_P, g_N, d_N. Depending on the cell type, only two of the traits have an effect on its phenotype. In the photosynthetic cell, trait g_P is the fraction of fixed carbon kept for the reproduction of the cell, the remainder, $1 - g_P$, is given out to neighbouring nitrogen fixing cells. The trait d_P defines the frequency with which photosynthetic cells differentiate into nitrogen fixing cells. In a nitrogen fixing cell, the traits g_N, d_N define similar behaviours as in the photosynthetic cell: g_N is the fraction of fixed nitrogen kept and d_N the fraction nitrogen fixing cells that differentiate. (b) The elementary steps in the model are cell death and cell division. Cells are chosen for division based on their fitness. When a cell divides, the offspring is inserted between the parent cell and a random neighbour. For cell death, cells are chosen at random without regards to their fitness. (c) Two different filament topologies were investigated. The connected topology, in which all cells remain connected after a cell death. And the broken chain topology, in which a cell death results in the separation of its neighbours. (d) The effects of interaction range were investigated by increasing the number of connections between the cells and their nearest neighbours.

such as cell size [161], the rate of biomass production, maintenance costs of the cell [162], or even regulatory effects. For example, such differences are observed in the case of grass leafs which grow slower when they contain larger amounts of cellulose and lignin [163].

In the model, every cell is characterised by four evolvable traits (g_P, d_P, g_N, d_N) which may have any value in the range $[0, 1]$ (Figure 4.1a). Of these four traits, two traits (g_P, d_P) affect only photosynthetic cells, while the other two (g_N, d_N) affect only nitrogen fixing cells. The traits g_P or g_N control how much of the resources produced by a cell are kept for its own growth, while the remaining fraction $1 - g_P$ or $1 - g_N$ is given away to neighbouring cells. The traits d_P or d_N control the fraction of offspring cells that differentiate into the other cell type. For example, if a photosynthetic cell has the trait $d_P = 1$ then all its offspring will differentiate into nitrogen fixing cells. The individuals in our simulations evolve through mutation. This can occur every time a cell reproduces at which time traits in the offspring may mutate with probability $\mu = 0.01$, changing by a random amount uniformly distributed in the range $[-0.1, 0.1]$.

Cell reproductive fitness is determined by growth rate in the model. Although we refer only to cell growth rates, the latter is synonymous with cell division rate in this case. A cell's growth rate depends on the amount of carbohydrates and fixed nitrogen produced and received from other cells. While the amount of resources received from other cells will depend on other factors such as the cell interaction topology, the interaction range and the traits of the other cells (Figure 4.1). Given these considerations, we define the fitnesses of a photosynthetic cell i and a nitrogen fixing cell j as

$$f_{P_i} = \alpha \, min(g_{P_i}, R_{N_i}) + f_{base} \tag{4.1}$$

$$f_{N_j} = \frac{g_{N_j} R_{C_j}}{2} + f_{base}. \tag{4.2}$$

The coefficient α expresses the difference in relative growth rates between the photosynthetic cells and the nitrogen fixing cells. The parameter f_{base} is a small constant that represents the base fitness and serves only to prevent the fitness from being zero. In our simulations we have used a value of $f_{base} = 0.001$. R_{N_i} and R_{C_j} are, respectively, the amounts of resources received by the photosynthetic cell i and the nitrogen fixing cell j from its neighbours and can be written as

$$R_{C_j} = \sum_{i}^{k_{P_j}} \frac{1 - g_{P_i}}{k_{N_i}} \tag{4.3}$$

$$R_{N_i} = \sum_{j}^{k_{N_i}} \frac{(1 - g_{N_j}) R_{C_j}}{k_{P_j}}. \tag{4.4}$$

The term k_{N_i} is the number of nitrogen fixing neighbours of the photosynthetic cell i, and k_{P_j} is the number of photosynthetic cells in the neighbourhood of the nitrogen fixing cell j.

To study the effects of differentiation costs we have modelled such costs as a reduction in the fitness of a differentiated cell by a fraction (C) such that the fitness of the cell becomes $f'_P = f_P(1-C)$. After the first time a cell is chosen for division, this cost is removed. Differentiation costs are considered because the process of differentiation is known to incur costs in higher order organisms [164]. Such costs can also be expected to exist in differentiating cells because differentiation requires a cell to degrade the proteins corresponding to its previous phenotype. The degradation of these proteins therefore incurs a cost of energy or materials. The existence of such costs is suggested by the fact that in terminally differentiating cyanobacteria, the cells which differentiate into nitrogen fixing heterocysts are the smaller ones resulting from asymmetric division of photosynthetic vegetative cells [88, 165, 166].

Fitness in our model is translated into a proportional probability that the cell will be chosen for reproduction every iteration. This probability is given by $P_i = f_i/f_T$ where f_i is the fitness of a cell and f_T is the sum over all the fitnesses of the cells in a population.

In this model, cells are arranged in linear chains. When a cell reproduces, a new cell with the same traits is inserted in the chain between its parent and a neighbour (Figure 4.1b). We investigate two filament topologies that result as a consequence of the type of cell death considered (Figure 4.1b and 4.1c). In the broken chain topology, the chain is broken in two parts when a cell chosen for death is removed, hence separating some of the neighbours of the removed cell. In the connected topology, a cell chosen for death is simply removed from the chain, with one of the neighbours taking the place of the removed cell. In addition, we study the effects of varying the number of interactions by increasing the interaction range (K) between cells (Figure 4.1d).

A notable characteristic of the model is that while in cyanobacteria the vegetative cells are the germline and the heterocysts are the somatic cells, this distinction is not fixed here but is instead an evolvable property which depends on the traits that control the cell type's reproduction and differentiation rate. This allows us to investigate the conditions under which different types of differentiation evolve.

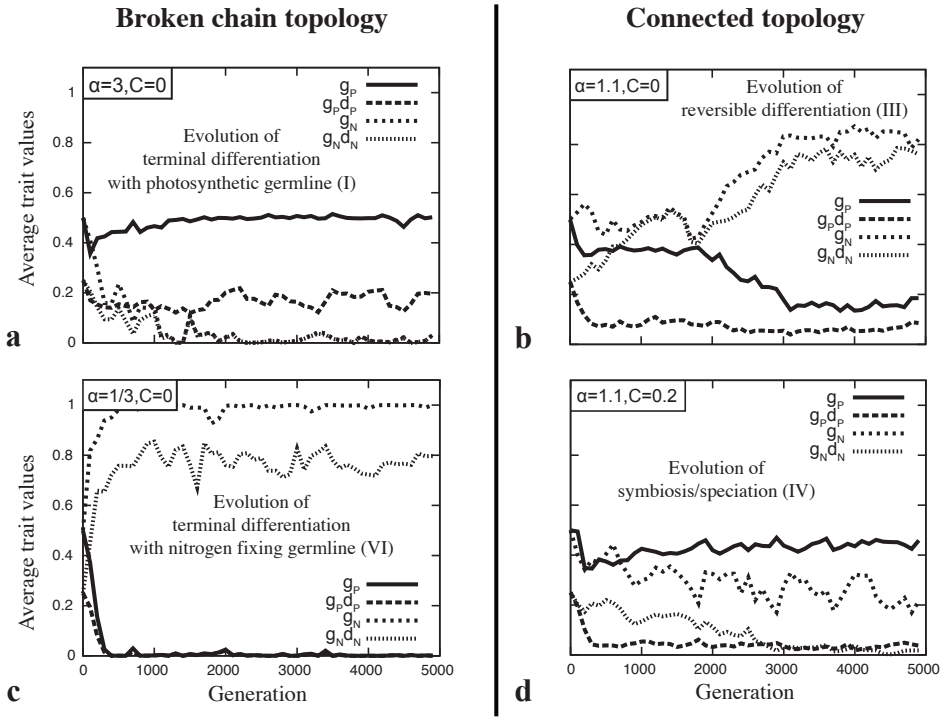

Figure 4.2: Examples of the evolution of the population trait averages (g_P, $g_P.d_P$, g_N, $g_N.d_N$) of 400 cells over 5000 generations under different conditions of relative growth rate α, filament topology, and differentiation costs C. Results shown in panels (a) and (c) correspond to the broken chain topology and differ only in the relative growth rates ($\alpha = 3$) and ($\alpha = 1/3$), respectively. Results shown in panels (b) and (d) correspond to the connected topology and differ only in the differentiation cost ($C = 0$) and ($C = 0.2$), respectively. All simulations were carried out with interaction range ($K = 22$).

4.4 Results

We analysed the evolution of the variable traits (g_P, d_P, g_N, d_N) in populations of 400 cells starting with a given set of initial values for all cells with the variable traits ($g_P = 0.5$, $d_P = 0.5$, $g_N = 0.5$, $d_N = 0.5$). Initially all cells were placed in a circular connected filament. Cells were randomly assigned as photosynthetic or nitrogen fixing with equal probability. The four panels in Figure 4.2 show examples of the evolution of the population average of each trait in four different conditions. Each generation corresponds to 400 cell deaths and divisions. Instead of the differentiation rates d_P and d_N, the products $g_P.d_P$ and $g_N.d_N$ are plotted because these express the effective rate of differentiation after cell division. In all simulations shown in Figure 4.2, it can be seen that the average variable traits evolve rapidly in the first generations until a point were they start fluctuating around a certain state that depends on the parameters of the simulation. Simulations using random initial variable traits, different population sizes and lower mutation rates did not affect the final states of the simulations. The parameters investigated are the relative growth rate α, differentiation cost C, filament topology, and interaction range K. Using the averages of the variable traits we classify the evolved developmental strategy of the population at each generation using Table 4.1. For the purpose of classification, we consider trait values bellow the threshold of 0.05 to be effectively 0. Figure 4.2a shows the evolution of the averages of variable traits (g_P, $g_P.d_P$, g_N, $g_N.d_N$) over 5000 generation of an evolving population using a broken chain topology, where photosynthetic cells have a relative grow rate three times faster ($\alpha = 3$) than nitrogen fixing cells, and with no differentiation costs ($C = 0$). We can see that in the final generation, photosynthetic cells keep half of the produced carbohydrates ($g_P = 0.5$) for their own cell growth and differentiate at a rate of ($g_P.d_P = 0.2$), while the nitrogen fixing cells do not keep any fixed nitrogen ($g_N \approx 0$) and therefore do not grow nor differentiate ($g_N.d_N \approx 0$). Using Table 4.1 we can classify this strategy as terminal differentiation with a photosynthetic germline (I). Figure 4.2c shows a simulation in the same conditions as in Figure 4.2a except that the photosynthetic cells grow three times more slowly ($\alpha = 1/3$). In this case we observe that the final strategy is terminal differentiation with a nitrogen fixing germline (VI). Figures 4.2b and 4.2d show simulations in the connected topology with slightly faster growing photosynthetic cells ($\alpha = 1.1$). In Figure 4.2b there are no differentiation costs ($C = 0$) and the final strategy corresponds to reversible differentiation (III). In Figure 4.2d there is a differentiation cost ($C = 0.2$) and the final strategy corresponds to the case of symbiosis (IV) (the different cell types evolve into separate lineages).

Next we investigate whether different developmental strategies may evolve in the same conditions. Figure 4.3 shows the plots of frequencies of the evolution of each developmental strategy when 50 stochastic simulations are carried out in the same

Developmental strategy	g_P	d_P	g_N	d_N
I. Terminal differentiation with photosynthetic germline	+	+	0	*
II. Terminal differentiation with photosynthetic germline with somatic division	+	+	+	0
III. Reversible differentiation	+	+	+	+
IV. Symbiotic/Speciation	+	0	+	0
V. Terminal differentiation with nitrogen fixing germline with somatic division	+	0	+	+
VI. Terminal differentiation with nitrogen fixing germline	0	*	+	+

Table 4.1: Developmental strategy classification based on variable traits (g_P,d_P,g_N,d_N). g_P and g_N correspond to the trait that determines the fraction of produced carbohydrates or fixed nitrogen kept for cell growth by the photosynthetic or the nitrogen fixing cell, respectively. d_P and d_N correspond to the traits that control the differentiation rate of photosynthetic and nitrogen fixing cells, respectively. The sign (+) indicates that the trait value has to be greater than zero in that strategy. The asterisk (*) indicates that the trait may have any value. For the purposes of classification, we considered trait values bellow the threshold of 0.05 to be effectively 0.

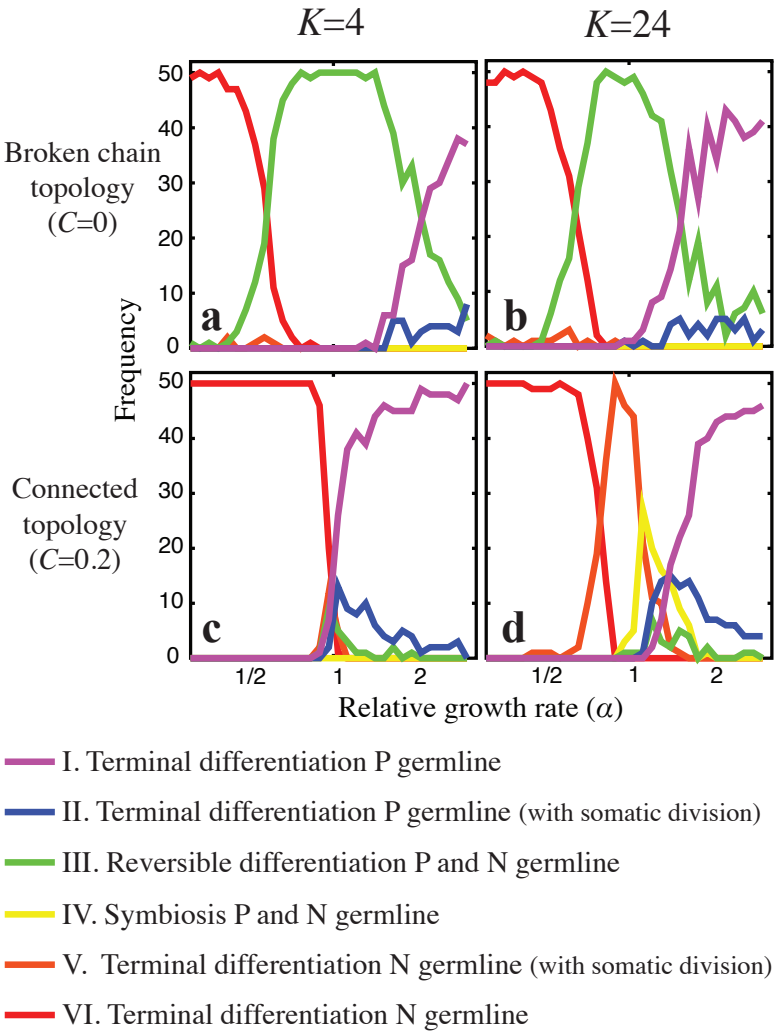

Figure 4.3: Frequency of evolved developmental strategies. The plots show the frequency of each strategy for varying relative growth rates α (50 simulations per α value). The frequency of each strategy is colour coded according to the key on the bottom. Two different cases are shown: (a,b) broken chain topology with no differentiation costs ($C = 0$), (c,d) connected topology with differentiation costs ($C = 0.2$). Each case is shown for two different interaction ranges ($K = 4, 24$) corresponding to the panels on the left, and right, respectively. Each simulation was performed with 400 cells over 5000 generations.

conditions. Each plot shows how the frequencies change with varying relative growth rate. The panels on the top (Figure 4.3a and 4.3b) show the results in the case of the broken chain topology with no differentiation costs ($C = 0$). And the plots on the bottom (Figure 4.3c and 4.3d) show the case of the connected topology with differentiation costs ($C = 0.2$). Each of filament topologies (Figure 4.3) were simulated with two different cell interaction range ($K = 4, 24$) which correspond to the plots on the left (Figures 4.3a and 4.3c), and right (Figures 4.3b and 4.3d), respectively.

In all the panels in Figure 4.3, generally only a single strategy is seen to evolve under a set of conditions with other strategies occurring only seldomly. However, some cases where two or more strategies evolve at appreciable frequencies can be observed. This is specially the case at points in which there is a transition in the most frequent strategy. For example, at $\alpha = 1/2$ in Figure 4.3a (broken chain topology, $K = 4$), a transition of the most frequently evolving strategies can be seen between terminal differentiation with nitrogen fixing germline (VI, red) and reversible differentiation (III, green). Another case is seen at α slightly larger than 1 in Figure 4.3d (connected topology, $K = 24$), where many strategies can be seen to evolve with some frequency.

An observation common to all the panels in Figure 4.3 is that at large differentials in growth rates ($\alpha << 1$ or $\alpha >> 1$), when one cell grows much faster than the other, terminal differentiation without somatic division evolves. Furthermore, it is the faster growing cell type that becomes the germline. Hence, at low relative growth rates ($\alpha << 1$), when nitrogen fixing cells are the faster growing, terminal differentiation with a nitrogen fixing germline (VI, red) is the most frequently evolved strategy. While at high relative growth rates ($\alpha >> 1$), when photosynthetic cells are the faster growing, terminal differentiation with photosynthetic germline (I, violet) is the most frequently evolved strategy.

The cases shown in Figure 4.3 and two other cases (the broken chain topology with differentiation costs $C = 0.2$ and the connected topology with no differentiation costs $C = 0$) are shown in Figure 4.5 for three different interaction ranges ($K = 4, 12, 24$).

To further examine the conditions which determine the most frequently evolved developmental strategies, we performed simulations for different relative growth rates (α) ranging from $\alpha = 1/3$ to $\alpha = 3$, interaction ranges (K) ranging from $K = 2$ to $K = 24$, using the two different filament topologies (broken chain and connected), and two values of differentiation costs ($C = 0$ and $C = 0.2$). The panels in Figure 4.4 show the most frequently evolved strategies, represented as colours classified in panel (a), for each combination of parameters α and K in simulations repeated 50 times. All cases shown confirm that terminal differentiation (I, violet and VI, red) evolves at the extremes of relative growth rate in which the

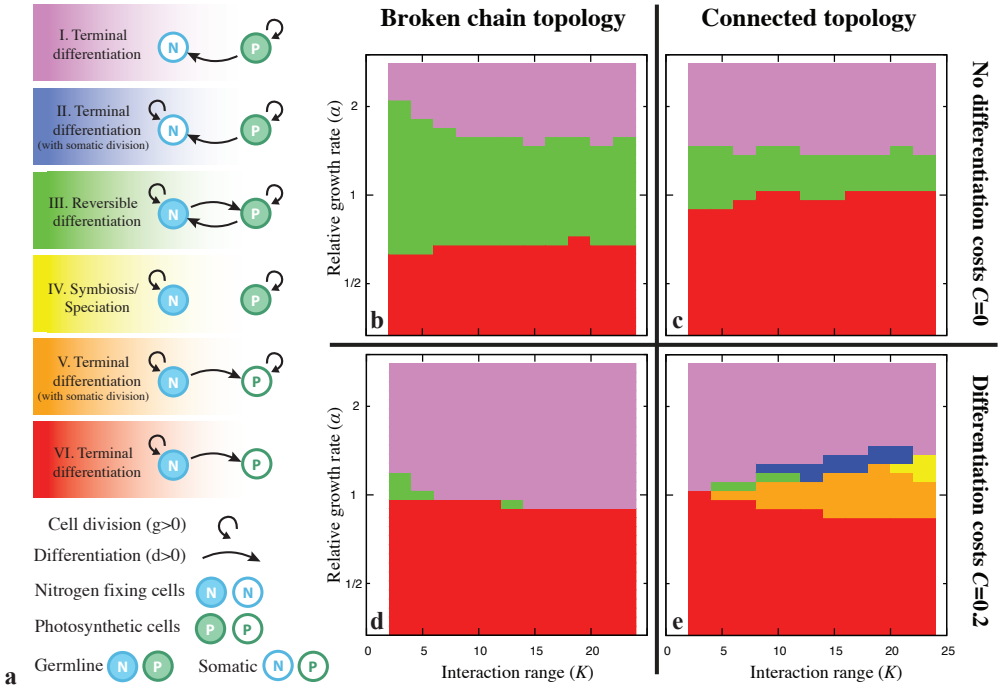

Figure 4.4: (a) Developmental strategies classified based on the final trait averages (g_P, d_P, g_N, d_N). The arrows that point from one cell type to itself represent investment in growth (g_P or g_N) while the arrows between cell types represent differentiation (d_P or d_N). Six possible developmental strategies exist: I. terminal differentiation with photosynthetic germline and non-dividing nitrogen fixing soma (violet), II. terminal differentiation with photosynthetic germline and dividing nitrogen fixing soma (blue), III. Reversible differentiation (green), IV. symbiosis (yellow), V. terminal differentiation with nitrogen fixing germline and dividing photosynthetic soma (orange), and VI. terminal differentiation with nitrogen fixing germline and non-dividing photosynthetic soma (red). The four panels on the right show the most frequently evolved developmental strategies depending on the cell interaction range (K) and the relative growth rate of photosynthetic cells (α). Panels (b) and (d) show the results simulated in the broken chain topology with no differentiation costs ($C = 0$) and with differentiation costs ($C = 0.2$), respectively. Panels (c) and (e) show the results in the connected topology with the same two differentiation costs. Simulations were repeated 50 times for each parameter combination, and the population size was 400. The colour represents the strategy found to evolve most frequently, with colour codes as in panel (a) (see also Table 4.1).

fastest growing cell becomes the germline.
In the broken chain topology, both with no differentiation costs $C = 0$ (Figure 4.4b) and with differentiation costs $C = 0.2$ (Figure 4.4d), only three developmental strategies can be observed in the conditions we examined. These are differentiation with a photosynthetic germline (I, violet), reversible differentiation (III, green), and terminal differentiation with a nitrogen fixing germline (VI, red). In both cases it can be seen that the main factor influencing the evolved developmental strategy is the relative growth rate (α), with little dependency on the interaction range of the cells (K). In Figure 4.4b, where no differentiation costs were included, fast growing photosynthetic cells ($\alpha > 2$) result in the evolution of terminal differentiation with photosynthetic cells as the germline (I, violet). Slow growing photosynthetic cells ($\alpha < 0.6$) also lead to the evolution of terminal differentiation, but in this case the nitrogen fixing cells become the germline (VI, red). For intermediate relative growth rates ($2 > \alpha > 0.6$), reversible differentiation (III, green) is the evolved strategy.
When a differentiation cost $C = 0.2$ is considered (Figure 4.4d), the range under which reversible differentiation (III, green) evolves is reduced to $1.2 > \alpha > 0.9$. Conversely, the range of α values under which terminal differentiation (I, violet and VI, red) evolves increases.
For the connected topology, with no differentiation costs $C = 0$ (Figure 4.4c) the result is qualitatively similar to the one observed for the broken chain topology with $C = 0$ (Figure 4.4b). In both cases only three strategies are observed to evolve most frequently, the two types of terminal differentiation without somatic division (I, violet and VI, red) and reversible differentiation (III, green).
Remarkably, when considering differentiation costs $C = 0.2$ (Figure 4.4e), all developmental strategies evolve in some range of conditions. Reversible differentiation (III, green) is reduced to a very narrow range of conditions with intermediate values of interaction ranges ($4 < K < 12$) and slightly faster growing photosynthetic cells ($\alpha \approx 1.1$). The range of conditions previously occupied by reversible differentiation (III, green) is replaced by terminal differentiation with somatic division (II, blue and V, orange) at shorter interaction ranges ($K < 20$), and symbiosis (IV, yellow) at longer interaction ranges ($K > 20$).

4.5 Discussion

4.5.1 Importance of differential growth rate

The results shown here establish a strong link between the relative growth rate of different cell types and the cell type that becomes the germline in a multicellular organism. The results in Figures 4.3 and 4.4 at high growth rate differences indi-

cate that when one cell type grows faster than the other, it evolves to become the germline in the organism. This result is found to be independent of the differentiation costs (C), filament topology, and interaction range (K) that characterise the organism. The reason can be explained intuitively by noting that an organism that requires both cell types will grow faster when the fastest growing cell type is the one that produces the other cell type as needed. Hence the faster growing cell types are the ones which remain pluripotent. For example this pattern is seen in plants, where cells in the apical meristems which generate shoots and roots consist of rapidly growing undifferentiated cells [167, 168]. Equivalently, one can interpret this as a situation in which cells that have a higher fitness at the individual level are the ones that become the germline.
When growth rates of the different cell types are comparable and $C = 0$, our model shows that reversible differentiation (III, green) evolves (Figures 4.4b and 4.4c). This corresponds to the case of differentiated cells that have the ability to de-differentiate into another cell type. Such cases are known to occur in many plants and in some animals capable of regeneration [84, 83].
The model indicates that although terminal differentiation is found to evolve in the widest range of conditions, reversible differentiation can evolve in conditions where the growth rates of different cell types are comparable. The latter can happen even in the absence of selection for the ability to regenerate or reproduce by fragmentation (Figures 4.4c,e).
It is important to note that large differences in cellular growth rates are a necessary but insufficient condition for a cell type to become the germline. The fast growth rate of a cell type must not harm the fitness of the organism as a whole, otherwise fast growing cells such as cancer cells would become the germline more often. Such an eventuality has occurred in only rare occasions [169, 170].

4.5.2 Role of filament topology and interaction range

Cell interaction affects developmental strategies in two ways. First, the broken chain topology increases the range of conditions under which reversible differentiation (III, green) evolves when compared with the connected topology (Figure 4.4b). The reason can be understood if we consider that reversible differentiation increases the survival of filaments in response to fragmentation. By ensuring that either cell type can produce the other cell type, the probability that a fragment will carry only non-differentiating cells is reduced. A similar argument can be made to explain why symbiosis (IV, yellow) does not evolve in the broken chain topologies under any conditions (Figures 4.4b and 4.4d). In these topologies, broken fragments never come into contact again, meaning that once a symbiotic pair within a filament is split, it will be condemned to death. Hence, such mutants can never become fixed in the population.

The effect of interaction range (K) is mainly seen in connected topologies. In this case, all possible developmental strategies evolve in at least one set of conditions (Figure 4.4c,e). For example, the symbiotic state (IV, yellow) that was not found in broken chain topologies, occurs in the connected topology when interaction ranges (K) are sufficiently high ($K > 20$) and there are differentiation costs ($C = 0.2$). In the case with differentiation costs, increasing the interaction range leads to a decrease in the range of relative growth rates under which terminal differentiation evolves, while the range for other strategies expands (Figure 4.4e). Higher interaction ranges ($24 <= K <= 40$) in the connected topology are shown in Figure 4.7. They lead to a slight increase in the range of relative growth rates in which symbiosis (IV, yellow) and terminal differentiation with a nitrogen fixing germline and somatic division (V, orange) occur.

It is well known that topologies with few interactions promote cooperative behaviour, while fully connected topologies, where all individuals interact with each other, result in the invasion of cheaters [171, 172]. This has already been shown to be the case in a model of cyanobacteria [157], in which populations of vegetative and heterocyst cells are driven to extinction in the fully connected case. Here, we have analysed topologies that are far from the fully connected case and where several forms of cooperation are stable.

Why a terminally differentiated cell forfeits its reproduction can in principle be explained by commonly accepted conditions that favour cooperation. Hence, the notion of inclusive fitness [159] may help explain why terminal differentiation arises instead of reversible differentiation. However, by varying the relative growth rate, the model shows that several developmental strategies such as reversible differentiation and symbiosis can evolve in the same filament topology and interaction range (Figure 4.4). These developmental strategies are neither altruistic nor selfish since both cell types can divide. Hence, the mapping of our present results to established concepts in social biology may require further work.

4.5.3 Correspondence to developmental strategies in cyanobacteria

Multicellular cyanobacteria have evolved several of the developmental strategies seen in this model. Terminally differentiating cyanobacteria such as *Anabaena* or *Nostoc* have filamentous forms composed of two different cell types: vegetative cells that are photosynthetic, divide and differentiate into the other cell type, and heterocyst cells, that fix nitrogen and are unable to divide. The latter can be distinguished by their larger size and thicker cell walls [88]. Our model provides clues to why heterocystous cyanobacteria form heterocysts by terminal differentiation without heterocyst division. An ad-hoc explanation that is frequently used is that

a heterocyst's thicker cell wall impedes it from undergoing cell division. However, our results provide an alternative explanation. In light of this model, a thicker cell wall corresponds to added costs and therefore a slower growth rate. Under this condition, the developmental strategy that maximises the organism's fitness is terminal differentiation without somatic division (I, violet and VI, red) (Figure 4.4d). This means that the reason why heterocysts do not divide is not necessarily due to mechanistic constraints, but rather a result of evolutionary constraints.
The only known example of potentially reversibly differentiated cyanobacteria are *Trichodesmium*. In species of this genus, different cell types are morphologically indistinguishable. However, differences at the level of expression of nitrogenase exist, and nitrogen fixation is shown to occur in distinct cells found across the filaments [92]. Although cells are differentiated in their expressed protein and function, both cell types maintain their ability to divide [173, 174]. While no direct experiment has shown that cells in *Trichodesmium* reversibly differentiate, the fact that the fraction of nitrogen fixing cells varies with daily rhythmicity, reaching a maximum of 24% during the day and a minimum of 5% before dawn, suggests that the nitrogen fixing cells may reversibly differentiate into photosynthetic cells [175]. In this case again, our results provide some insights as to why cells that are specialised in nitrogen fixation (therefore similar to heterocysts) are not terminally differentiated, but are still capable of dividing and of reverting back to a photosynthetic phenotype. Because both cell types are structurally similar, they can be expected to have similar growth rates. The results shown in Figures 4.3a and 4.4b predict that reversible differentiation (III, green) should be the most frequently evolved developmental strategy in this case.
So far, no known examples of multicellular cyanobacteria exist in which terminally differentiating nitrogen fixing cells (heterocysts) are capable of cell division (II, blue). While this can simply reflect our incomplete knowledge, our results suggest that such developmental strategies are evolutionarily unstable (Figure 4.4b-e).

4.5.4 Symbiosis/speciation

The finding that symbiosis evolves in a connected topology under several conditions of relative cell growth rate and differentiation costs points to some interesting evolutionary possibilities. One is that some organisms may have speciated as a result of changing conditions that initially selected for terminal or reversible differentiation, but later changed to favour a symbiotic state. Potential support for this idea comes from a recently sequenced cyanobacterium named UCYN-A that is closely related to a member species of the genus *Cyanothece* [176]. *Cyanothece* are unicellular circadian cyanobacteria capable of photosynthesis and nitrogen fixation by temporally separating the two processes. The newly sequenced relative of *Cyanothece* lacks the genes necessary to perform photosynthesis found in *Cyan-*

othece species [176]. Instead, it has only the genes necessary for nitrogen fixation. Because it is unable to perform photosynthesis, it is dependent on obtaining its carbohydrates from the environment or from other organisms. This suggests that a scenario in which cyanobacteria speciate into symbiotic or interacting collectives is possible. In effect, chloroplasts, which are endosymbionts that descended from cyanobacteria, are a likely endpoint of such a scenario. In the latter case, chloroplasts provide the host plant with fixed carbon while the plant is the intermediary that provides fixed nitrogen.
Plants have never evolved the ability to fix nitrogen. They absorb it from the environment or rely instead on symbiotic diazotrophic bacteria such as the cyanobacterium *Nostoc* to fix nitrogen in exchange for carbohydrates produced by the plant through photosynthesis. The vascular system of plants changes the topology of cell interactions from a chain to a connected topology with high interaction ranges, which suggests that all the photosynthetic cells in the plant are able to exchange nutrients with the nitrogen fixing cyanobacteria in the roots of the plant. Our results show that in such conditions (Figures 4.3d and 4.4e), a symbiotic relationship (IV, yellow) where the nitrogen fixing cells evolve independently from the photosynthetic cells is the most frequent evolved strategy. The range of α values in which symbiosis evolves is seen to increase with higher differentiation costs and interaction ranges (Figures 4.6 and 4.7, respectively). These results suggest that the symbiotic relationship between plants and cyanobacteria may be evolutionarily more stable than the alternative scenario, in which plants would fix their own nitrogen.

4.5.5 Generality of the model

While this model draws inspiration from differentiated cyanobacteria, the results found here may apply to a wider range of biological systems. In essence, the model describes the evolution of a multicellular organism or population with two types of individuals that produce different resources, but require both to reproduce. Hence, these individuals need to interact and exchange resources. By considering the exchange of fitness benefits as a form of resource exchange, a cell type in an organism that serves a structural function can also be analysed using such a model. One assumption we have made that may not apply to other systems is that the nitrogen fixing cells are only able to fix nitrogen provided they obtain some carbohydrates from photosynthetic cells. This results in an asymmetry in the model because photosynthetic cells do not require fixed nitrogen to perform photosynthesis. We show in Figure 4.8 that the results presented here do not qualitatively change when we modify the model to enable nitrogen fixing cells to fix nitrogen independently of the carbohydrates received. This suggests that the results are robust with regard to the specific type of resource exchange considered.

Hence the framework presented here may be sufficiently general to apply to other systems.

4.5.6 Conclusion

We have shown that the topology of interactions, the interaction range and the relative growth rate between cells in an organism all play a role in the type of differentiation that evolves. However, the relative growth rate, and hence differential rates of cell division, was found to play the most important role. Not only is differential cell growth the main factor determining the type of differentiation that evolves, but it also determines the cell type that becomes the germline. Differential growth rates can be influenced by many factors such as physiological and environmental conditions, or be internally regulated. If our results serve as an indication of what can be expected in organisms with more than two cell types, then it implies that the evolution of different forms of differentiation in multicellular organisms is subject to constraints based on the maximum relative growth rates between cell types.

4.6 Supplementary Information

4.6.1 Frequency of evolved developmental strategies for different filament topologies, interaction ranges, and differentiation costs

Figure 4.5 shows the plots of frequencies of the evolution of each developmental strategy when 50 stochastic simulations are carried out in the same conditions. Each plot shows how the frequencies change with varying relative growth rate. The panels in row (a) in Figure 4.5 show the results in the case of the broken chain topology with no differentiation costs ($C = 0$). The panels in row (b) in Figure 4.5 show the broken chain topology with differentiation costs ($C = 0.2$). The panels in row (c) in Figure 4.5 show the case of the connected topology with differentiation costs ($C = 0.2$). Each of these four cases (Figure 4.5) were simulated with a varying cell interaction range ($K = 4, 12, 24$) which correspond to the plots on the left, middle, and right columns, respectively.

4.6.2 Higher differentiation costs (C) and interaction ranges (K) favor symbiosis in the connected topology

In Figure 4.6 we examined the effect of increasing the differentiation costs ($C = 0.3$) in the connected topology. Here we compare these results to the case where the differentiation costs are lower ($C = 0.2$), as shown in Figure 4e. First, we observe that higher differentiation costs reduce the range of conditions under which terminal differentiation without somatic division (violet and red) evolves. Instead, the symbiotic strategy (yellow), and terminal differentiation with somatic division, where the heterocyst is the germline (orange) evolve under a broader range of conditions. As a result, the dependency on interaction range becomes stronger, where increasing interaction ranges lead to a higher probability that one of these strategies evolves.

In Figure 4.7 we show the evolved developmental strategies at longer interaction ranges ($K = 24$ to $K = 40$) in the connected topology, in the cases of no differentiation costs ($C = 0$) and modest differentiation costs ($C = 0.2$). Longer interaction ranges in the case of no differentiation costs (Figure 4.7a) do not change the results seen at shorter interaction ranges (Figure 4c). In the case with differentiation costs, longer interaction ranges (Figure 4.7b) increase slightly the range of relative growth rates under which symbiosis (yellow) and terminal differentiation with somatic division and a heterocyst germline (orange) occur when compared to shorter interaction ranges (Figure 4e).

4.6.3 Qualitatively similar results are found in the symmetric model

In the model presented, the fitness of a photosynthetic cell and a nitrogen fixing cell is described by equations (3) and (4) in the main text. These equations describe a model that assumes that a nitrogen fixing cell can only perform its function when it is supplied with carbohydrates. One can also consider a different case where the ability to fix nitrogen is independent of the supply of carbohydrates. The latter assumption leads to symmetric fitness functions for the photosynthetic and nitrogen fixing cells where:

$$f_{P_i} = \alpha\, min(g_{P_i}, R_{N_i}) + f_{base} \tag{4.5}$$

$$f_{N_j} = min(g_{N_j}, R_{C_j}) + f_{base}. \tag{4.6}$$

Because the photosynthetic cells are no longer the only source of energy in the system, both cells become equal partners needing the products from each other. In Figure 4.8 we show the developmental strategies that evolve in this symmetric model, in comparable conditions to the asymmetric model (Figure 4 in the main text). The results are qualitatively similar. However, in this case the evolved strategies symmetric around the relative cell growth rate (α). In addition, we find that the symbiotic strategy is found to be restricted to a much narrower set of conditions characterised by high differentiation costs ($C = 0.6$).

4.7 Supplementary figures

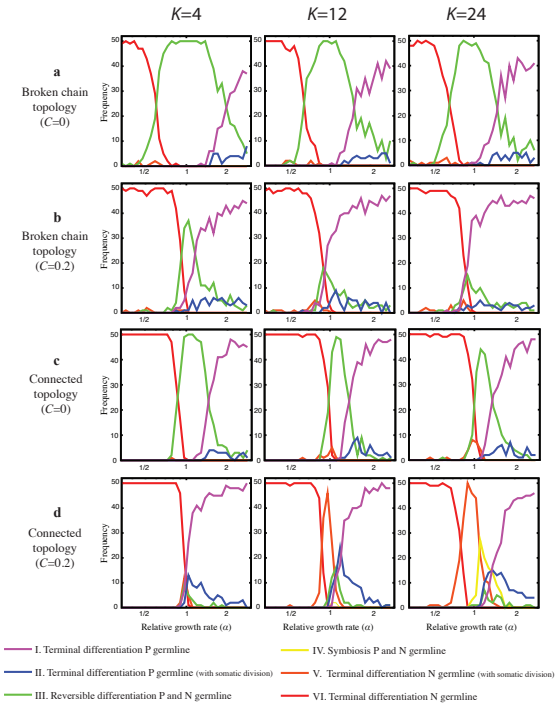

Figure 4.5: Frequency of evolved developmental strategies. The plots show the frequency of each strategy with varying relative growth rates α (50 simulations per α value). The frequency of each strategy is colour coded according to the key on the bottom. Four different cases are shown: (row a) broken chain topology with no differentiation costs ($C = 0$), (row b) broken chain topology with differentiation costs ($C = 0.2$), (row c) connected topology with no differentiation costs ($C = 0$), and (row d) connected topology with differentiation costs ($C = 0.2$). The plots in the three different columns correspond to different interaction ranges ($K = 4, 12, 24$), as shown above each column. Simulations were performed with 400 cells over 5000 generations.

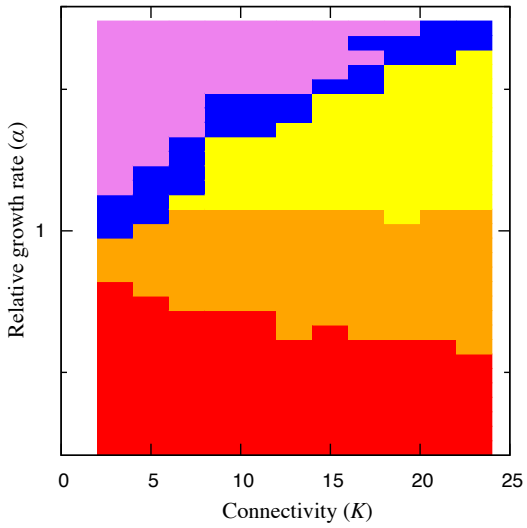

Figure 4.6: Most frequently evolved developmental strategies in the connected topology. The simulations were performed with varying cell interaction range K and photosynthetic cell relative growth rate α with differentiation cost ($C = 0.3$). Simulations were repeated 50 times for each parameter combination and the population size was 400. The colour represents the most frequently evolved strategy coded according to Figure 4a in the main text.

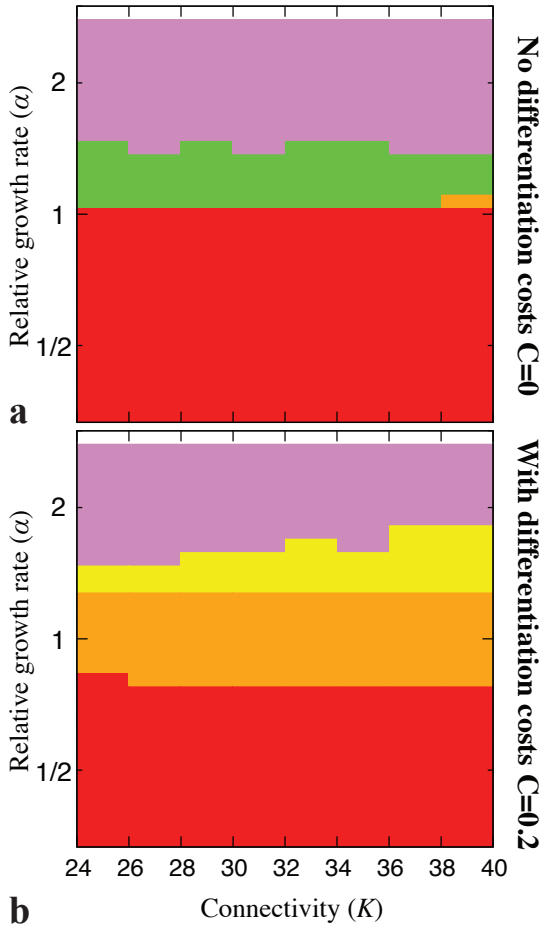

Figure 4.7: Most frequently evolved developmental strategies in the connected topology with high interaction ranges K between ($K = 24$) and ($K = 40$). The two panels show the results of the simulations (a) with no differentiation costs ($C = 0$) and (b) with differentiation costs ($C = 0.6$). Simulations were repeated 50 times for each parameter combination, and the population size was 400. The colour represents the most frequently evolved strategy coded according to Figure 4a in the main text.

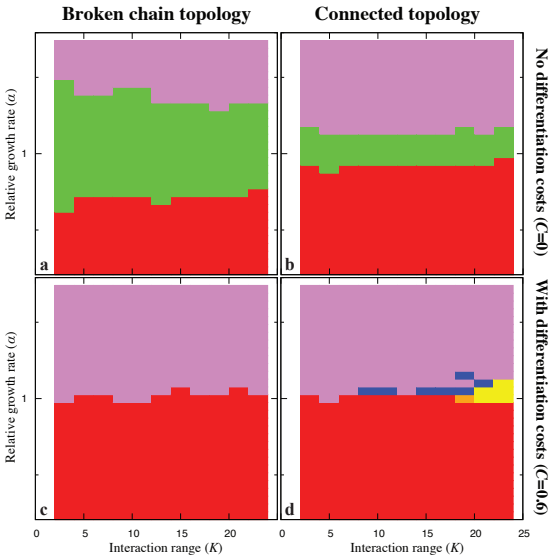

Figure 4.8: Most frequently evolved developmental strategies in simulations where different cell types have symmetric fitnesses. Panels (a) and (c) show the results of the broken chain topology. Panels (b) and (d) show the results in the connected chain topology. The simulations were performed with varying cell interaction ranges K and photosynthetic cell relative growth rates α, (a,b) with no differentiation costs ($C = 0$) and (c,d) with differentiation costs ($C = 0.6$). Simulations were repeated 50 times for each parameter combination, with population sizes of 400. The colour represents the most frequently evolved strategy coded according to Figure 4a in the main text.

5 Conclusion

In my thesis I have modeled the evolution of two different aspects of microbes. The first aspect of microbial evolution I study in my thesis concerns the evolution of metabolism. I develop a model in which the evolution of metabolic networks occurs through horizontal gene transfer and loss of function mutations or gene deletion. Using this model I show that the existence of constant-phenotype genotype networks are found also in metabolism. These genotype networks facilitate the encounter of novel potentially beneficial phenotypes by allowing organisms to evolve while maintaining a constant phenotype. Because such genotype networks have been found in other biological systems, these results strengthen the idea that constant-phenotype genotype networks are a general property of evolvable systems. I also find that not all properties found in other biological systems are found in the genotype-phenotype map of metabolic networks. In the genotype-phenotype map of RNA sequences and structures, it was found that more robust phenotypes are correlated with a higher number of encountered novel phenotypes. However, this is not observed in metabolism, where metabolic networks of intermediate robustness encounter the highest number of novel phenotypes.
The second aspect of microbial evolution that I study in my thesis concerns the evolution of differentiation in a multicellular organism or in a population of cells that exchange resources. Specifically, I develop a model that describes a population of interacting cells that require both fixed carbon and fixed nitrogen to grow and reproduce but in which each cell is unable to produce both resources at the same time. I find that differences between the two cell types that affect their relative cell growth rate, the physiological conditions in which the cells interact, and differentiation costs favor the evolution of different developmental strategies. This is the first model capable of explaining the different developmental strategies observed in extant multicellular cyanobacterial species. These results may also shed some insight on how development evolves in higher organisms and also explain why some cell types become terminally differentiated while others are reversibly differentiated.

Bibliography

[1] Canfield D, Desmarais D. Biogeochemical cycles of carbon, sulfur, and free oxygen in a microbial mat. Geochimica et Cosmochimica Acta. 1993 Aug;57(16):3971–3984.

[2] Falkowski PG. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. Nature. 1997 May;387(6630):272–275.

[3] Falkowski PG. The role of phytoplankton photosynthesis in global biogeochemical cycles. Photosynthesis Research. 1994 Mar;39(3):235–258.

[4] Amann RI, Ludwig W, Schleifer KH. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiological reviews. 1995 Mar;59(1):143–69.

[5] Grice EA, Kong HH, Conlan S, Deming CB, Davis J, Young AC, et al. Topographical and temporal diversity of the human skin microbiome. Science. 2009 May;324(5931):1190–2.

[6] Kroes I. Bacterial diversity within the human subgingival crevice. Proceedings of the National Academy of Sciences. 1999 Dec;96(25):14547–14552.

[7] Paster BJ, Boches SK, Galvin JL, Ericson RE, Lau CN, Levanos VA, et al. Bacterial diversity in human subgingival plaque. Journal of bacteriology. 2001 Jun;183(12):3770–83.

[8] Hattori M, Taylor TD. The Human Intestinal Microbiome: A New Frontier of Human Biology. DNA Research. 2009 Feb;16(1):1–12.

[9] Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, et al. Environmental genome shotgun sequencing of the Sargasso Sea. Science. 2004 Apr;304(5667):66–74.

[10] Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, et al. Comparative metagenomics of microbial communities. Science. 2005 Apr;308(5721):554–7.

[11] Zuckerkandl E, Pauling L. Evolutionary divergence and convergence in proteins. Evolving Genes and Proteins. 1965;p. 97–165.

[12] Kimura M. The Neutral Theory of Molecular Evolution. Cambridge, UK: Cambridge University Press; 1983.

[13] Fitch WM, Margoliash E. Construction of Phylogenetic Trees. Science. 1967 Jan;155(3760):279–284.

[14] Fox GE, Stackebrandt E, Hespell RB, Gibson J, Maniloff J, Dyer TA, et al. The phylogeny of prokaryotes. Science. 1980 Jul;209(4455):457–63.

[15] Boto L. Horizontal gene transfer in evolution: facts and challenges. Proceedings of the Royal Society B: Biological Sciences. 2009 Oct;277(1683):819–827.

[16] Brown JR. Ancient horizontal gene transfer. Nature reviews Genetics. 2003 Feb;4(2):121–32.

[17] Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000 May;405(6784):299–304.

[18] Koonin EV, Makarova KS, Aravind L. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. Annual Review of Microbiology. 2003 Nov;55(1):709–742.

[19] Thomas CM. The Horizontal Gene Pool: Bacterial Plasmids and Gene Spread. Amsterdam: Harwood Academic Publishers; 2000.

[20] Thomas CM, Nielsen KM. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nature reviews Microbiology. 2005 Sep;3(9):711–21.

[21] Lorenz MG, Wackernagel W. Bacterial gene transfer by natural genetic transformation in the environment. Microbiological reviews. 1994 Sep;58(3):563–602.

[22] Paget E, Simonet P. On the track of natural transformation in soil. FEMS Microbiology Ecology. 1994;.

[23] Gregory TR, Hebert PDN. The Modulation of DNA Content: Proximate Causes and Ultimate Consequences. Genome Res. 1999;9(4):317–324.

[24] Koonin EV, Wolf YI. Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic acids research. 2008 Dec;36(21):6688–719.

[25] Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. Trends in Genetics. 2001 Oct;17(10):589–596.

[26] Lynch M, Conery JS. The origins of genome complexity. Science (New York, NY). 2003 Nov;302(5649):1401–4.

[27] Andersson S, Kurland CG. Reductive evolution of resident genomes. Trends in Microbiology. 1998 Jul;6(7):263–268.

[28] Blanc G, Ogata H, Robert C, Audic S, Suhre K, Vestris G, et al. Reductive genome evolution from the mother of Rickettsia. PLoS genetics. 2007 Jan;3(1):e14.

[29] Gray MW, Burger G, Lang BF. Mitochondrial Evolution. Science. 1999 Mar;283(5407):1476–1481.

[30] Wernegreen JJ. Genome evolution in bacterial endosymbionts of insects. Nature reviews Genetics. 2002 Nov;3(11):850–61.

[31] Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JCD, Andersson DI. Bacterial genome size reduction by experimental evolution. Proceedings of the National Academy of Sciences of the United States of America. 2005 Aug;102(34):12112–6.

[32] Huynen MA, Dandekar T, Bork P. Variation and evolution of the citric-acid cycle: a genomic perspective. Trends in microbiology. 1999 Jul;7(7):281–91.

[33] Romano A. Evolution of carbohydrate metabolic pathways. Research in Microbiology. 1996 Sep;147(6-7):448–455.

[34] Horowitz NH. On the Evolution of Biochemical Syntheses. Proceedings of the National Academy of Sciences of the United States of America. 1945 Jun;31(6):153–7.

[35] Horowitz NH. Part I. In: Bryson V, Vogel HJ, editors. Evolving Genes and Proteins. New York: Academic Press; 1965. p. 15–23.

[36] Jensen RA. Enzyme recruitment in evolution of new function. Annual review of microbiology. 1976 Jan;30:409–25.

[37] Copley SD. Enzymes with extra talents: moonlighting functions and catalytic promiscuity. Current opinion in chemical biology. 2003 Apr;7(2):265–72.

[38] Khersonsky O, Roodveldt C, Tawfik DS. Enzyme promiscuity: evolutionary and mechanistic aspects. Current opinion in chemical biology. 2006 Oct;10(5):498–508.

[39] Hult K, Berglund P. Enzyme promiscuity: mechanism and applications. Trends in biotechnology. 2007 May;25(5):231–8.

[40] Teichmann SA, Rison SC, Thornton JM, Riley M, Gough J, Chothia C. The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. Journal of molecular biology. 2001 Aug;311(4):693–708.

[41] Light S, Kraulis P. Network analysis of metabolic enzyme evolution in Escherichia coli. BMC Bioinformatics. 2004;5(15).

[42] Edwards JS, Palsson BO. Systems Properties of the Haemophilus influenzae Rd Metabolic Genotype. Journal of Biological Chemistry. 1999 Jun;274(25):17410–17416.

[43] Edwards JS, Palsson BO. Metabolic flux balance analysis and the in silico analysis of Escherichia coli K-12 gene deletions. BMC Bioinformatics. 2000;1(1):1.

[44] Martins S. A kinetic model for the glucose/glycine Maillard reaction pathways. Food Chemistry. 2005 Apr;90(1-2):257–269.

[45] de Noronha Pissara P, Nielsen J, Bazin MJ. Pathway kinetics and metabolic control analysis of a high-yielding strain ofPenicillium chrysogenum during fed batch cultivations. Biotechnology and Bioengineering. 1996 Jul;51(2):168–176.

[46] Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. Current Opinion in Biotechnology. 2003 Oct;14(5):491–496.

[47] Edwards JS, Palsson BO. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. Proceedings of the National Academy of Sciences of the United States of America. 2000 May;97(10):5528–33.

[48] Gerdes SY, Scholle MD, Campbell JW, Balazsi G, Ravasz E, Daugherty MD, et al. Experimental Determination and System Level Analysis of Essential Genes in Escherichia coli MG1655. Journal of Bacteriology. 2003 Sep;185(19):5673–5684.

[49] Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. Proceedings of the National Academy of Sciences. 2002 Nov;99(23):15112–15117.

[50] Kanehisa M, Goto S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Research. 2000 Jan;28(1):27–30.

[51] Edwards JS, Ibarra RU, Palsson BO. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nature biotechnology. 2001 Feb;19(2):125–30.

[52] Ibarra RU, Edwards JS, Palsson BO. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature. 2002 Nov;420(6912):186–9.

[53] Ingraham JL, Maaloe O, Neidhardt FC. Growth of the Bacterial Cell. Sunderland, MA: Sinauer Associates Inc.; 1983.

[54] Varma A, Palsson BO. Metabolic Capabilities of Escherichia coli II. Optimal Growth Patterns. Journal of Theoretical Biology. 1993 Dec;165(4):503–522.

[55] Murty KG. Linear Programming. New York: John Wiley & Sons; 1983.

[56] Edwards JS, Palsson BO. Robustness analysis of the Escherichia coli metabolic network. Biotechnology progress. 2000;16(6):927–39.

[57] Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using Escherichia coli. Nature biotechnology. 2008 Jun;26(6):659–67.

[58] Oh YK, Palsson BO, Park SM, Schilling CH, Mahadevan R. Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. The Journal of biological chemistry. 2007 Sep;282(39):28791–9.

[59] Segura D, Mahadevan R, Juárez K, Lovley DR. Computational and experimental analysis of redundancy in the central metabolism of Geobacter sulfurreducens. PLoS computational biology. 2008 Feb;4(2):e36.

[60] Alberch P. From genes to phenotype: dynamical systems and evolvability. Genetica. 1991 May;84(1):5–11.

[61] Lipman DJ, Wilbur WJ. Modelling neutral and selective evolution of protein folding. Proceedings Biological sciences / The Royal Society. 1991 Jul;245(1312):7–11.

[62] Schuster P, Fontana W, Stadler PF, Hofacker IL. From sequences to shapes and back: a case study in RNA secondary structures. Proceedings Biological sciences / The Royal Society. 1994 Mar;255(1344):279–84.

[63] Ciliberti S, Martin OC, Wagner A. Innovation and robustness in complex regulatory gene networks. Proceedings of the National Academy of Sciences of the United States of America. 2007 Aug;104(34):13591–6.

[64] Babajide A, Hofacker IL, Sippl MJ, Stadler PF. Neutral networks in protein space: a computational study based on knowledge-based potentials of mean force. Folding & design. 1997;2(5):261–269.

[65] Amitai G, Gupta RD, Tawfik DS. Latent evolutionary potentials under the neutral mutational drift of an enzyme. HFSP Journal. 2007 May;1(1):67–78.

[66] Bershtein S, Goldin K, Tawfik DS. Intense neutral drifts yield robust and evolvable consensus proteins. Journal of Molecular Biology. 2008 Jun;379(5):1029–1044.

[67] Bloom JD, Romero PA, Lu Z, Arnold FH. Neutral genetic drift can alter promiscuous protein functions, potentially aiding functional evolution. Biology direct. 2007 Jun;2.

[68] Kitano H. Biological robustness. Nature reviews Genetics. 2004 Nov;5(11):826–37.

[69] de Visser JAGM, Hermisson J, Wagner GP, Meyers LA, Bagheri-Chaichian H, Blanchard JL, et al. Perspective: Evolution and Detection of Genetic Robustness. Evolution. 2003;57(9):1959 – 1972.

[70] van Nimwegen E, Crutchfield JP, Huynen M. Neutral evolution of mutational robustness. Proceedings of the National Academy of Sciences. 1999 Aug;96(17):9716–9720.

[71] Montville R, Froissart R, Remold SK, Tenaillon O, Turner PE. Evolution of mutational robustness in an RNA virus. PLoS biology. 2005 Nov;3(11):e381.

[72] Wagner A, Stadler PF. Viral RNA and evolved mutational robustness. Journal of Experimental Zoology. 1999 Aug;285(2):119–127.

[73] Wagner A. Distributed robustness versus redundancy as causes of mutational robustness. BioEssays : news and reviews in molecular, cellular and developmental biology. 2005 Feb;27(2):176–88.

[74] Wagner A. Robustness and evolvability: a paradox resolved. Proceedings Biological sciences / The Royal Society. 2008 Jan;275(1630):91–100.

[75] Lenski RE, Barrick JE, Ofria C. Balancing robustness and evolvability. PLoS biology. 2006 Dec;4(12):e428.

[76] Wagner A. OPINION Neutralism and selectionism: a network-based reconciliation. Nature Reviews Genetics. 2008 Dec;9(12):965–974.

[77] Maynard Smith J, Szathmáry E. The major transitions in evolution. Oxford: Freeman; 1995.

[78] Michod RE. Evolution of individuality during the transition from unicellular to multicellular life. Proceedings of the National Academy of Sciences of the United States of America. 2007 May;104 Suppl(suppl_1):8613–8.

[79] Rainey PB, Rainey K. Evolution of cooperation and conflict in experimental bacterial populations. Nature. 2003 Sep;425(6953):72–4.

[80] Grosberg RK, Strathmann RR. The Evolution of Multicellularity: A Minor Major Transition? Annual Review of Ecology, Evolution, and Systematics. 2007;38(1):621–654.

[81] Pfeiffer T, Bonhoeffer S. An evolutionary scenario for the transition to undifferentiated multicellularity. Proceedings of the National Academy of Sciences of the United States of America. 2003 Feb;100(3):1095–8.

[82] Birnbaum KD, Sánchez Alvarado A. Slicing across kingdoms: regeneration in plants and animals. Cell. 2008 Feb;132(4):697–710.

[83] Carnevali C. Regeneration in Echinoderms: repair, regrowth, cloning. Invertebrate Survival Journal. 2006;3(1):64–76.

[84] Sánchez Alvarado A, Tsonis PA. Bridging the regeneration gap: genetic insights from diverse animal models. Nature reviews Genetics. 2006 Nov;7(11):873–84.

[85] Gallon J. The oxygen sensitivity of nitrogenase: a problem for biochemists and micro-organisms. Trends in Biochemical Sciences. 1981;6:19–23.

[86] Golden SS, Ishiura M, Johnson CH, Kondo T. Cyanobacterial Circadian Rhythms. Annual Review of Plant Physiology and Plant Molecular Biology. 1997;48(1):327–354.

[87] Adams DG. Multicellularity in cyanobacteria. In: Mohan S, Dow C, Cole JA, editors. Prokaryotic structure and function: a new perspective. Society for General Microbiology symposium. Cambridge, UK: Cambridge University Pres; 1992. p. 341–384.

[88] Adams DG, Duggan PS. Tansley Review No. 107 Heterocyst and Akinete Differentiation in Cyanobacteria. New Phytologist. 1999;144(1):3 – 33.

[89] Berman-Frank I, Lundgren P, Chen YB, Küpper H, Kolber Z, Bergman B, et al. Segregation of nitrogen fixation and oxygenic photosynthesis in the marine cyanobacterium Trichodesmium. Science. 2001 Nov;294(5546):1534–7.

[90] Mylona P, Pawlowski K, Bisseling T. Symbiotic Nitrogen Fixation. The Plant cell. 1995 Jul;7(7):869–885.

[91] Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. Generic Assignments, Strain Histories and Properties of Pure Cultures of Cyanobacteria. Microbiology. 1979 Mar;111(1):1–61.

[92] Lin S, Henze S, Lundgren P, Bergman B, Carpenter EJ. Whole-Cell Immunolocalization of Nitrogenase in Marine Diazotrophic Cyanobacteria, Trichodesmium spp. Appl Envir Microbiol. 1998;64(8):3052–3058.

[93] Meeks JC, Elhai J. Regulation of Cellular Differentiation in Filamentous Cyanobacteria in Free-Living and Plant-Associated Symbiotic Growth States. Microbiology and Molecular Biology Reviews. 2002 Mar;66(1):94–121.

[94] Li WH. Molecular Evolution. Massachusetts: Sinauer Associates Inc.; 1997.

[95] Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome biology. 2003 Jan;4(9):R54.

[96] Förster J, Famili I, Fu P, Palsson BO, Nielsen J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome research. 2003 Feb;13(2):244–53.

[97] Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. Nature reviews Microbiology. 2004 Nov;2(11):886–97.

[98] Wagner A. Robustness and evolvability in living systems. Princeton, NJ: Princeton University Press; 2005.

[99] Pál C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nature genetics. 2005 Dec;37(12):1372–5.

[100] Harrison R, Papp B, Pál C, Oliver SG, Delneri D. Plasticity of genetic interactions in metabolic networks of yeast. Proceedings of the National Academy of Sciences of the United States of America. 2007 Mar;104(7):2307–12.

[101] Price ND, Papin JA, Palsson BO. Determination of redundancy and systems properties of the metabolic network of Helicobacter pylori using genome-scale extreme pathway analysis. Genome research. 2002 May;12(5):760–9.

[102] Stelling J, Klamt S, Bettenbrock K, Schuster S, Gilles ED. Metabolic network structure determines key aspects of functionality and regulation. Nature. 2002 Nov;420(6912):190–3.

[103] Almaas E, Oltvai ZN, Barabási AL. The activity reaction core and plasticity of metabolic networks. PLoS computational biology. 2005 Dec;1(7):e68.

[104] Gerdes SY, Scholle MD, Campbell JW, Balázsi G, Ravasz E, Daugherty MD, et al. Experimental determination and system level analysis of essential genes in Escherichia coli MG1655. Journal of bacteriology. 2003 Oct;185(19):5673–84.

[105] Heinrich R, Schuster S. The regulation of cellular systems. New York: Chapman and Hall; 1996.

[106] Pál C, Papp B, Lercher MJ, Csermely P, Oliver SG, Hurst LD. Chance and necessity in the evolution of minimal metabolic networks. Nature. 2006 Mar;440(7084):667–670.

[107] Segrè D, Deluna A, Church GM, Kishony R. Modular epistasis in yeast metabolism. Nature genetics. 2005 Jan;37(1):77–83.

[108] Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, et al. Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. Molecular systems biology. 2006 Jan;2:2006.0008.

[109] Alper H, Miyaoku K, Stephanopoulos G. Construction of lycopene-overproducing E. coli strains by combining systematic and combinatorial gene knockout targets. Nature biotechnology. 2005 May;23(5):612–6.

[110] Al Zaid Siddiquee K, Arauzo-Bravo MJ, Shimizu K. Metabolic flux analysis of pykF gene knockout Escherichia coli based on 13C-labeling experiments together with measurements of enzyme activities and intracellular

metabolite concentrations. Applied microbiology and biotechnology. 2004 Jan;63(4):407–17.

[111] Yang C, Hua Q, Baba T, Mori H, Shimizu K. Analysis of Escherichia coli anaplerotic metabolism and its regulation mechanisms from the metabolic responses to altered dilution rates and phosphoenolpyruvate carboxykinase knockout. Biotechnology and bioengineering. 2003 Oct;84(2):129–44.

[112] Joyce AR, Reed JL, White A, Edwards R, Osterman A, Baba T, et al. Experimental and computational assessment of conditionally essential genes in Escherichia coli. Journal of bacteriology. 2006 Dec;188(23):8259–71.

[113] Wunderlich Z, Mirny LA. Using the topology of metabolic networks to predict viability of mutant strains. Biophysical journal. 2006 Sep;91(6):2304–11.

[114] Blank LM, Kuepfer L, Sauer U. Large-scale 13C-flux analysis reveals mechanistic principles of metabolic network robustness to null mutations in yeast. Genome biology. 2005 Jan;6(6):R49.

[115] Motter AE, Gulbahce N, Almaas E, Barabási AL. Predicting synthetic rescues in metabolic networks. Molecular systems biology. 2008 Jan;4:168.

[116] Papp B, Pál C, Hurst LD. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. Nature. 2004 Jun;429(6992):661–4.

[117] Schilling CH, Palsson BO. Assessment of the metabolic capabilities of Haemophilus influenzae Rd through a genome-scale pathway analysis. Journal of theoretical biology. 2000 Apr;203(3):249–83.

[118] Vitkup D, Kharchenko P, Wagner A. Influence of metabolic network structure and function on enzyme evolution. Genome biology. 2006 Jan;7(5):R39.

[119] Goto S, Nishioka T, Kanehisa M. LIGAND: chemical database for enzyme reactions. Bioinformatics (Oxford, England). 1998 Jan;14(7):591–9.

[120] Schuster S, Dandekar T, Fell DA. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. Trends in biotechnology. 1999 Feb;17(2):53–60.

[121] Schilling CH, Schuster S, Palsson BO, Heinrich R. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. Biotechnology progress. 1999;15(3):296–303.

[122] Kuepfer L, Sauer U, Blank LM. Metabolic functions of duplicate genes in Saccharomyces cerevisiae. Genome research. 2005 Oct;15(10):1421–30.

[123] DeLuna A, Vetsigian K, Shoresh N, Hegreness M, Colón-González M, Chao S, et al. Exposing the fitness contribution of duplicated genes. Nature genetics. 2008 May;40(5):676–81.

[124] Wang Z, Zhang J. Abundant indispensable redundancies in cellular metabolic networks. Genome biology and evolution. 2009 Jan;2009:23–33.

[125] Myllykallio H, Leduc D, Filee J, Liebl U. Life without dihydrofolate reductase FolA. Trends in microbiology. 2003 May;11(5):220–3.

[126] Li H, Helling R, Tang C, Wingreen N. Emergence of preferred structures in a simple model of protein folding. Science (New York, NY). 1996 Aug;273(5275):666–9.

[127] Chan HS, Bornberg-Bauer E. Perspectives on protein evolution from simple exact models. Applied bioinformatics. 2002 Jan;1(3):121–44.

[128] Xia Y, Levitt M. Roles of mutation and recombination in the evolution of protein thermodynamics. Proceedings of the National Academy of Sciences of the United States of America. 2002 Aug;99(16):10382–7.

[129] Goodman M, Pedwaydon J, Czelusniak J, Suzuki T, Gotoh T, Moens L, et al. An evolutionary tree for invertebrate globin sequences. Journal of molecular evolution. 1988 Jan;27(3):236–49.

[130] Rost B. Protein structures sustain evolutionary drift. Folding & design. 1997 Jan;2(3):S19–24.

[131] Bloom JD, Labthavikul ST, Otey CR, Arnold FH. Protein stability promotes evolvability. Proceedings of the National Academy of Sciences of the United States of America. 2006 Apr;103(15):5869–74.

[132] Aharoni A, Gaidukov L, Khersonsky O, McQ Gould S, Roodveldt C, Tawfik DS. The 'evolvability' of promiscuous protein functions. Nature genetics. 2005 Jan;37(1):73–6.

[133] Boucher I, Parrot M, Gaudreau H, Champagne CP, Vadeboncoeur C, Moineau S. Novel food-grade plasmid vector based on melibiose fermentation for the genetic engineering of Lactococcus lactis. Applied and environmental microbiology. 2002 Dec;68(12):6152–61.

[134] Reed JL, Palsson BO. Genome-scale in silico models of E. coli have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. Genome research. 2004 Sep;14(9):1797–805.

[135] Fischer E, Sauer U. Large-scale in vivo flux analysis shows rigidity and suboptimal performance of Bacillus subtilis metabolism. Nature genetics. 2005 Jun;37(6):636–40.

[136] Almaas E, Kovács B, Vicsek T, Oltvai ZN, Barabási AL. Global organization of metabolic fluxes in the bacterium Escherichia coli. Nature. 2004 Feb;427(6977):839–43.

[137] Pramanik J, Keasling JD. Stoichiometric model ofEscherichia coli metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. Biotechnology and Bioengineering. 1997 Nov;56(4):398–421.

[138] Neidhardt FC, Ingraham JL, Schaechter M. Physiology of the bacterial cell. Sunderland, MA: Sinauer Associates Inc.; 1990.

[139] Matias Rodrigues JaF, Wagner A. Evolutionary plasticity and innovations in complex metabolic reaction networks. PLoS computational biology. 2009 Dec;5(12):e1000613.

[140] Samal A, Matias Rodrigues JaF, Jost J, Martin OC, Wagner A. Genotype networks in metabolic reaction spaces. BMC systems biology. 2010 Jan;4:30.

[141] Ferrada E, Wagner A. Protein robustness promotes evolutionary innovations on large evolutionary time-scales. Proceedings Biological sciences / The Royal Society. 2008 Jul;275(1643):1595–602.

[142] Huxtable RJ. Biochemistry of Sulfur. New York: Plenum Press; 1986.

[143] Sekowska A, Kung HF, Danchin A. Sulfur metabolism in Escherichia coli and related bacteria: facts and fiction. Journal of molecular microbiology and biotechnology. 2000 Apr;2(2):145–77.

[144] Hartl DL, Clark AG. Principles of population genetics. 4th ed. Sinauer Associates; 2007.

[145] Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, et al. Systematic screen for human disease genes in yeast. Nature genetics. 2002 Aug;31(4):400–4.

[146] Kondrashov AS, Sunyaev S, Kondrashov FA. Dobzhansky-Muller incompatibilities in protein evolution. Proceedings of the National Academy of Sciences of the United States of America. 2002 Nov;99(23):14878–83.

[147] Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. Nature. 2010 Mar;464(7286):279–82.

[148] Kirschner M. Evolvability. Proceedings of the National Academy of Sciences. 1998 Jul;95(15):8420–8427.

[149] Ferrada E, Wagner A. Evolutionary innovations and the organization of protein functions in genotype space. (submitted). 2010;.

[150] Schultes EA. One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. Science. 2000 Jul;289(5478):448–452.

[151] IUBMB. Enzyme Nomenclature. San Diego, California: Academic Press; 1992.

[152] Ewens WJ. Mathematical Population Genetics. Berlin: Springer; 1979.

[153] Gurdon JB. The developmental capacity of nuclei taken from intestinal epithelium cells of feeding tadpoles. Journal of embryology and experimental morphology. 1962 Dec;10:622–40.

[154] Eggan K, Baldwin K, Tackett M, Osborne J, Gogos J, Chess A, et al. Mice cloned from olfactory sensory neurons. Nature. 2004 Mar;428(6978):44–9.

[155] Campbell KH, McWhir J, Ritchie WA, Wilmut I. Sheep cloned by nuclear transfer from a cultured cell line. Nature. 1996 Mar;380(6569):64–6.

[156] Buss LW. Evolution, Development, and the Units of Selection. Proceedings of the National Academy of Sciences. 1983 Mar;80(5):1387–1391.

[157] Rossetti V, Schirrmeister BE, Bernasconi MV, Bagheri HC. The evolutionary path to terminal differentiation and division of labor in cyanobacteria. Journal of theoretical biology. 2010 Jan;262(1):23–34.

[158] Gavrilets S. Rapid transition towards the Division of Labor via evolution of developmental plasticity. PLoS computational biology. 2010 Jan;6(6):e1000805.

[159] Hamilton W. The genetical evolution of social behaviour. I. Journal of Theoretical Biology. 1964 Jul;7(1):1–16.

[160] Vrede K, Heldal M, Norland S, Bratbak G. Elemental Composition (C, N, P) and Cell Volume of Exponentially Growing and Nutrient-Limited Bacterioplankton. Applied and Environmental Microbiology. 2002 Jun;68(6):2965–2971.

[161] Tyson CB, Lord PG, Wheals AE. Dependency of size of Saccharomyces cerevisiae cells on growth rate. J Bacteriol. 1979;138(1):92–98.

[162] Pirt SJ. The Maintenance Energy of Bacteria in Growing Cultures. Proceedings of the Royal Society of London Series B, Biological Sciences. 1965;163(991):224 – 231.

[163] Arendonk JJCM, Poorter H. The chemical composition and anatomical structure of leaves of grass species differing in relative growth rate. Plant, Cell and Environment. 1994 Aug;17(8):963–970.

[164] DeWitt T, Sih A, Wilson DS. Costs and limits of phenotypic plasticity. Trends in Ecology & Evolution. 1998 Feb;13(2):77–81.

[165] Mitchison GJ, Wilcox M. Rule governing Cell Division in Anabaena. Nature. 1972 Sep;239(5367):110–111.

[166] Wilcox M, Mitchison GJ, Smith RJ. Pattern formation in the Blue-Green Alga, Anabaena: I. Basic Mechanisms. J Cell Sci. 1973;12(3):707–723.

[167] Medford JI. Vegetative Apical Meristems. The Plant cell. 1992 Sep;4(9):1029–1039.

[168] Kwiatkowska D. Flowering and apical meristem growth dynamics. Journal of experimental botany. 2008 Jan;59(2):187–201.

[169] Pearse AM, Swift K. Allograft theory: transmission of devil facial-tumour disease. Nature. 2006 Feb;439(7076):549.

[170] Murgia C, Pritchard JK, Kim SY, Fassati A, Weiss RA. Clonal origin and evolution of a transmissible cancer. Cell. 2006 Aug;126(3):477–87.

[171] Ohtsuki H, Hauert C, Lieberman E, Nowak MA. A simple rule for the evolution of cooperation on graphs and social networks. Nature. 2006 May;441(7092):502–5.

[172] Grafen A. An inclusive fitness analysis of altruism on a cyclical network. Journal of evolutionary biology. 2007 Nov;20(6):2278–83.

[173] El-Shehawy R, Lugomela C, Ernst A, Bergman B. Diurnal expression of hetR and diazocyte development in the filamentous non-heterocystous cyanobacterium Trichodesmium erythraeum. Microbiology. 2003 May;149(5):1139–1146.

[174] Fredriksson C, Bergman B. Ultrastructural characterisation of cells specialised for nitrogen fixation in a non-heterocystous cyanobacterium,Trichodesmium spp. Protoplasma. 1997 Mar;197(1-2):76–85.

[175] Fredriksson C, Bergman B. Nitrogenase quantity varies diurnally in a subset of cells within colonies of the non-heterocystous cyanobacteria Trichodesmium spp. Microbiology. 1995 Oct;141(10):2471–2478.

[176] Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, et al. Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. Nature. 2010 Mar;464(7285):90–4.

Curriculum vitae

Surname: MATIAS RODRIGUES
First Name: João Frederico
Date of Birth: 6th December 1980
Birthplace: Portugal

Education:
1998 Highschool, Liceu de Macau, Macau (Portugal)
1998-2004 Licenciatura in Physics, Universidade de Lisboa, Portugal
2007-2010 Dr. sc. nat., Universität Zürich, Switzerland